Cost-effective load testing of WebRTC applications

Francisco Gortázar*, Micael Gallego, Michel Maes-Bermejo, Iván Chicano, Carlos Santos

Escuela Técnica Superior de Ingeniería Informática, Universidad Rey Juan Carlos, Spain

{*francisco.gortazar, micael.gallego, michel.maes, ivan.chicano, carlos.santos*}*@urjc.es*

Abstract

Background: Video conference applications and systems implementing the WebRTC W3C standard are becoming more popular and demanded year after year, and load testing them is of paramount importance to ensure they can cope with demand. However, this is an expensive activity, usually involving browsers to emulate users. *Goal*: to propose alternative strategies for load testing WebRTC services, and to study performance and costs of those strategies when compared with traditional ones. *Method*: (a) Exploring the limits of existing and novel strategies for load testing WebRTC services from a single machine. (b) Comparing the common strategy of using browsers with the best of our proposed strategies in terms of cost in a stress testing scenario. *Results*: We observed that, using identical machines, our proposed strategies are able to emulate more users than traditional strategies. We also found a huge saving in expenditure for stress testing, as our strategy suppose a saving of 93% with respect to usual browser-based strategies. We also found there are almost no differences between the traditional strategies considered. *Conclusions*: We provide details on scalability of different load testing strategies in terms of users emulated, as well as CPU and memory used. We could reduce the expenditure of stress tests of WebRTC applications.

Keywords: testing, load testing, stress testing, WebRTC

*Corresponding author

1. Introduction

With the rise of remote jobs, businesses transiting online, or online courses, video conferencing has become an integral part of our lives. Video conference applications and systems are becoming more popular and demanded year after year. On top of this, the video conferencing market is experiencing high growth during the coronavirus outbreak. Whether this situation will persist after the coronavirus crisis or not is still something we don't know, nevertheless, the demand is pushing companies in the video conferencing market into a situation in which load testing their solutions is of paramount importance to ensure solutions can cope with demand.

Among the different technologies for real time video communication over the Internet, the World Wide Web's (W3C) Web Real-Time Communications (WebRTC) standard [1] is a popular one. Video conference tools like BigBlueButton [2], Jitsi [3], Whereby [4], among others all rely on the WebRTC standard to enable video calls for a set of users.

WebRTC is a set of protocols and APIs that provides web browsers and mobile applications with Real-Time Communications (RTC) capabilities over peer-to-peer connections. It was conceived to allow connecting browsers without intermediate helpers or services, but in practice this P2P model falls short when trying to create more complex applications. For this reason, in most cases a central media server is required. Conceptually, a WebRTC media server is just a multimedia middleware where media traffic passes through when moving from source(s) to destination(s). Media servers are capable of processing incoming media streams and offer different outcomes, such as:

- Group Communications: Distributing among several receivers the media stream that one peer generates, i.e. acting as a Selective Forwarding Unit ("SFU").
- Mixing: Transforming several incoming streams into one single composite stream, i.e. acting as a Multipoint Conferencing Units ("MCU").

- Transcoding: On-the-fly adaptation of codecs and formats between incompatible clients.

- Recording: Storing in a persistent way the media exchanged among peers.

In WebRTC, users of a group call (from now on session) are usually browsers exchanging real-time video and audio through a WebRTC media server. The number of users can vary, and the media server is usually able to handle many of such sessions. Testing group calls is usually an expensive scenario. Test cases are usually executed using several browsers to impersonate real users by exchanging preconfigured video and audio in real-time through a media server. Web browsers are heavy resource consumers, and usually a single browser impersonates a single user thus limiting scalability and a better use of resources. The challenges of testing video conference systems have been discussed in the past in the literature [5, 6, 7, 8].

In this paper: a) we propose two new testing strategies that avoid the need of a browser, and significantly reduces the expenditure of testing WebRTC based video conference applications; b) we perform a comparative study of different testing strategies for real-time group communications at scale using the WebRTC standard. In the study, we used the OpenVidu WebRTC platform [9], which is open source, and experiments were conducted in the cloud using Amazon Web Services [1]. We paid attention to benefits and drawbacks of the scrutinized strategies, the performance in terms of number of users per session, the use of resources, and the total expenditure of the tests.

We decided to compare traditional WebRTC testing strategies with our new proposals with two aims:

(1) To measure the scalability of each testing strategy within a single AWS machine (instance in AWS jargon) in terms of number of users that can be impersonated from that machine, by answering the following research questions:

- RQ_{1a}: What's the strategy that is able to emulate more users into ses-

[1] `https://aws.amazon.com/`

sions? Is this consistent with the session size?

- RQ_{1b}: How do each of the strategies perform in terms of CPU and memory with respect to the number of users?

(2) To compare the costs of our most promising testing strategy with traditional browser-based strategies in a stress test scenario, where multiple machines are needed to impersonate users:

- RQ_2: Are there any cost savings for using our proposals with respect to using browsers when doing stress testing of a WebRTC platform?

With the first set of research questions we try to understand the scalability and limits of each of the strategies in a single machine, and the correlation of the strategy with the use of resources. This information is used to answer the final research question, by conducting a stress test scenario with the most promising testing strategy. In this scenario, several testing machines are used, up to the limits defined in the first set of research questions, to explore the capacity of a WebRTC server in terms of users.

The rest of the paper is structured as follows: Section 2 discusses previous research in WebRTC testing (including stress or load testing of WebRTC services). In Section 3 we introduce the OpenVidu platform. Traditional testing strategies, and our new proposals for WebRTC load testing are presented in Section 4. We describe the methodology used in the study in Section 5. Results of applying the methodology are reported in Section 6, including a discussion on the threats to validity. Finally, Section 7 draws conclusions and presents further research.

2. Background

2.1. Verification and Validation (V&V)

Verification and Validation is the set of techniques used to assess software products and services. Software testing is a part of V&V, consisting of observing a sample of the executions of a piece of code using a subset of possible inputs and settings (test cases), evaluating the output from which a verdict is given [10].

Testing and quality assurance of web applications has the challenge of having to test heterogeneous applications [11]. According to Di Lucca and Fasolino [12], the main testing activities for non-functional requirements that a Web application is usually required to accomplish are: performance testing, load testing, stress testing, compatibility testing, usability testing, accessibility testing and security testing. In the remaining of the section strategies and tools from the literature (including gray literature) focused on generic testing, as well as load and stress testing, are presented.

2.2. Generic WebRTC testing tools

Among the WebRTC testing tools, Selenium [2] is one of the most widely used, as was reported previously by Garcia et al. [5]. Selenium enables programmatic managing of browsers using different programming languages, such as Java, Python, PHP or JavaScript. As most browsers currently implement the WebRTC stack, any WebRTC-enabled Selenium managed browser can be used to impersonate users for WebRTC testing. However, using browsers implies a considerable expenditure of computational resources. To avoid investment in infrastructure, users can resort to commercial providers of browsers such as Saucelabs [3], BrowserStack [4] and Nightwatch.js [5], in a pay-per-use approach.

The Kurento Testing Framework (KTF) [13] provides a set of browser-based tools for testing WebRTC applications. It allows to automatically evaluate functional parameters such as media events or color detection (for instance, to build oracles based on video synchronization), it collects performance statistics, and it is able to extract quality of experience metrics (by evaluating audio quality). This testing framework is a part of Kurento [14], a WebRTC platform, and it again uses Selenium for the connection with the browsers.

[2] `https://www.selenium.dev/`
[3] https://saucelabs.com/
[4] https://www.browserstack.com/
[5] https://nightwatchjs.org/

ElasTest [6] is a comprehensive open source platform that aims to significantly improve the efficiency and effectiveness of testing complex systems. It is a generic tool, not exclusively focused on WebRTC testing, although it includes interesting features for WebRTC testing. One of its strengths is observability, ElasTest platform provides a complete monitoring system for the software under test. When confronted with WebRTC testing, ElasTest, in addition to provide Selenium controlled browsers, it automatically obtains specific metrics of the WebRTC connection that browsers make available for inspection (the WebRTC stats defined within the WebRTC W3C standard).

KITE [6] is another framework for WebRTC peer-to-peer test automation supported and managed by companies actively involved in the development of the WebRTC standard. It is supported by other technologies such as Selenium and browser vendors. It allows interoperability testing between different browsers and operating systems. The main purpose of the project is to detect the level of compliance of each browser vendor with the WebRTC standard.

2.3. WebRTC stress testing tools

Some well-known load testing tools for web application such as Apache JMeter, Artillery or Gatling cannot be used for WebRTC testing, as they do not use a browser or any implementation of the WebRTC stack. Therefore, these tools are out of scope.

However, there are other tools in the literature aimed at stress testing a WebRTC media server. WebRTCBench [15] is a WebRTC benchmark which measures performance on peer-to-peer communication, allowing to evaluate the performance across different platform and devices, including the most popular browsers for desktop and Android platforms. The benchmark is limited to two peers, and therefore cannot be used to analyze the performance of media servers at scale, where many peers are connected simultaneously arranged into different media sessions.

[6] https://elastest.io/

In [16], the authors introduced JAttack, a WebRTC stress testing tool for performance analysis of the WebRTC media server Janus. This tool uses a modified Janus media server to generate multiple WebRTC connections in order to stress the Janus media server [17]. Although the authors performed some experiments stressing a Janus WebRTC media server, they only reported CPU usage of the testing tool and the media server. Furthermore, both the load testing tool (the modified Janus server) and the Janus media server were running together in the same machine, which raises some concerns about the validity of the results they got. Additionally, they did not perform a comparison and in depth analysis of the costs compared to traditional browser-based approaches. Finally, JAttack hasn't been released to the public.

The main commercial option is testRTC [18], which includes a wide variety of solutions for WebRTC application testing, including stress testing where network conditions can be configured to build realistic scenarios. As far as the authors can tell, testRTC prices are not publicly available, so cost analysis can not be performed.

2.4. Testing in the cloud

Cloud computing can provide users cost-effective and flexible scalable computing power and services. According to [19], the main benefits of cloud-based testing are: reduced costs of computing resources by leveraging them to clouds (using virtualized resources and shared infrastructure), existing on-demand test services for large-scale and effective real-time online validation and easily leverage scalable cloud system infrastructure to test and evaluate system performance and scalability.

Bertolino et al. [20] did a systematic review on cloud testing. They divided their study into testing in the cloud (leveraging the elasticity of cloud for testing), testing of the cloud (testing cloud applications) and testing of the cloud in the cloud (mixing both approaches). The authors conclude that testing in the cloud has indeed received much attention, being test execution the most prominent usage of cloud providers in the context of the different testing activities.

In [21], the authors tested the scalability of 4 WebRTC media servers (Medooze, Jitsi, Janus, and OpenVidu/Kurento), giving details of the infrastructure used. They used AWS Elastic Compute Cloud (EC2) service, separating the media server and the web clients on different instances. Specifically, the authors used the c4.4xlarge (8 vCPUs, 30 GB RAM) instance type for the media servers and the c4.xlarge (4 vCPUs, 7.5 GB RAM) instance type for the clients. The paper focuses mainly on the number of users supported by each media server, using KITE as the WebRTC testing platform. No analysis of costs was included in the work.

3. WebRTC Testing Strategies

Before introducing the different WebRTC testing strategies considered in the study, we present an overview of the architecture of WebRTC services and applications. In this Section, we first sketch how WebRTC application works and the services involved, and then we resort to detail how they are tested. Finally, we present our proposals for cost-effective load testing of WebRTC applications.

3.1. Anatomy of a WebRTC application

A WebRTC application comprises several components: 1) the application itself, running in the browser; 2) a library that enables the application to embed videoconference into the page; and 3) a media server that receives and routes streams of audio and video to all users in a session. The library in 2) makes transparent for the application how the WebRTC protocols are used, thus easing the embedding of video into the app. For instance, the library instructs the browser to fetch video and audio through the corresponding devices (camera and microphone) and it sends the streams to a media server through the WebRTC API.

Usually, media servers are able to handle multiple sessions running in parallel. For each session, a user in the session (connected through a browser) sends audio and video streams to the media server, which in turn, sends the video and audio streams that it receives from all other users in the same session to the user. Given a set of users in a session $N, |N| = n$, each user $x \in N$ sends 2 streams and receives $2 * (n-1)$ streams from the other $n-1$ users, $y \in N, y \neq x$ in the session. Depending on how the application is implemented, this might impose a limit on the number of users that a browser is able to handle for the session. For instance, Whereby limits to 50 the number of users in a session.

At the media server, the number of streams to be managed grows with the number of users in a session. For a session with 2 users, the media server receives 4 streams (2 per user), and it sends 2 streams to each user. With 3 users, the media server receives 6 streams, and it sends 4 streams to each user. This means $4 * 3 = 12$ streams to be sent in total to 3 different users. Therefore, the number of streams s to be sent depends quadratically on the number of users: $s = 2n * (n - 1) = 2n^2 - 2n$.

It is therefore of paramount importance to properly plan the size of the machine hosting and running the media server, as it will have to route many streams to different users in different sessions. In order to do so, proper stress tests need to be performed on the media server. As it was mentioned in Section 2, traditional approaches are based on web browsers to impersonate users in media sessions. This strategy works well, but it has a considerable cost. It is worth noticing that this stress testing might be needed for many scenarios, with different number of users in each session.

3.2. OpenVidu WebRTC platform

In order to study testing strategies for WebRTC we need a WebRTC application. OpenVidu [9] is an open source platform to enable embedding video conferencing capabilities into modern web and mobile applications. It provides both the browser library and the media server (the Kurento Media Server, KMS). In this section, we describe the OpenVidu architecture and how we leverage it to perform load tests in order to find the limits of the platform. We will conduct our comparative study using the OpenVidu platform and the OpenVidu Loadtest Tool (OVLT).

The OpenVidu platform comprises three modules as shown in Figure 1:

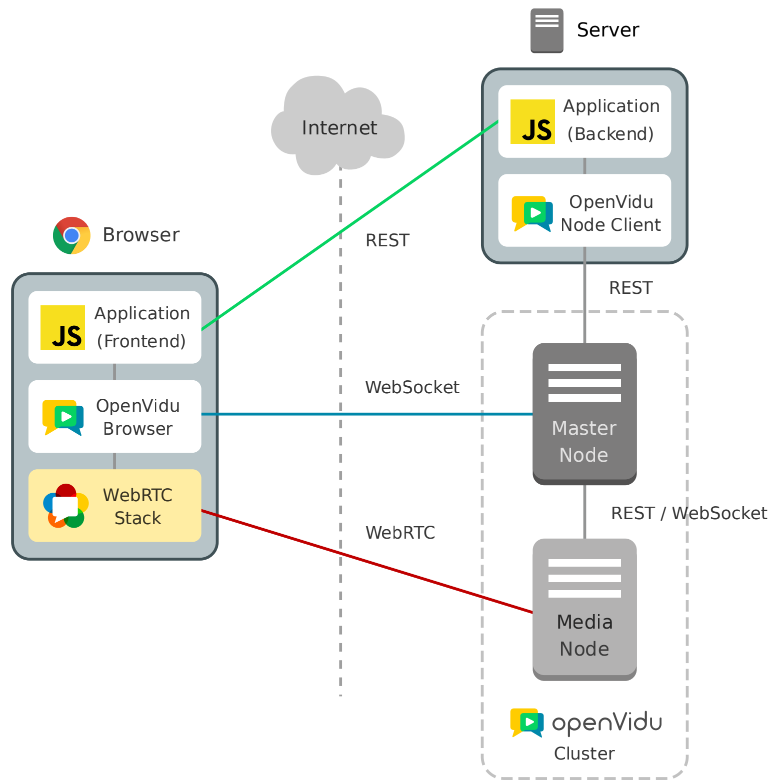

Figure 1: OpenVidu architecture. The OpenVidu platform consists of a master node and one or more media nodes. An application using the OpenVidu platform consists of a backend and a frontend. The openvidu-browser library is used by the frontend of the application to make use of OpenVidu WebRTC capabilities. The backend is usually used to provide security. Upon client authentication (green line), negotiation between the openvidu-browser library and the master node (blue line) results in a media node selection with whom the media will be exchanged with the help of the browser's implementation of the WebRTC stack (red line).

- **openvidu-browser**: A JavaScript library for the browser. It allows developers to embed video in their own applications. It is responsible for the control plane and leverages to the WebRTC stack provided by the browser for the media plane.
- **OpenVidu master node**: Executes an application that controls one or more media nodes. It takes care of the control plane, and it manages

sessions, forwarding events and messages to clients and distributing the load across the available media nodes.

- **OpenVidu media node**: Executes the media server responsible for the media plane. OpenVidu currently uses the Kurento Media Server, an open source implementation of an WebRTC media server. OpenVidu platform can use several media nodes. They are the actual bottleneck of the OpenVidu platform and their number and size determine its capacity: more media nodes means more concurrent sessions possible, which results in an increased spending.

OpenVidu provides a basic application, but developers can build their own application using the OpenVidu APIs. A WebRTC application based on OpenVidu consists on a backend and a frontend. In the backend, authentication and authorization of users is managed. The frontend requests session joins for a user through the backend, and upon authorization, a token is returned to the frontend. This token is used by openvidu-browser to connect to the master node asking for a specific session to join, and the master node selects which media node to use for the session. Thus, the master node is in charge of the control plane, managing where each session is hosted (i.e., in which media node). When the media node is selected, openvidu-browser uses the browser WebRTC API to negotiate how to exchange media between the browser and the media server. The media server and the browser WebRTC API are in charge of the media plane.

Notice that all the users of a session must all be connected to the same media node. That is, a session is fully contained into a single media node. A media node is able to handle several sessions concurrently. The specific number of sessions that a media node is able to handle depends on the number of users connected to the session (the session size) and is somehow difficult and cost ineffective to calculate. Hence, we are interested in researching new cost-effective strategies for finding the limits of a media node for different session sizes that might be more cost-effective and result in less expenditure.

3.3. OpenVidu load testing

Most testing efforts on the WebRTC arena have been focused on testing WebRTC compliance on the different browsers. However, little attention has been paid on easing the task of load testing WebRTC applications, with a few exceptions [16, 21]. Most testing approaches for these applications are still based on user impersonation through browsers, thus emulating a video conference session by starting several web browsers that are connected into a single session through a media server. This approach has several advantages: it is possible to record the video conference session in each browser and apply state-of-the-art Quality of Experience (QoE) algorithms to automatically determine quality, or even record the browser window to gain better insight on how the application is performing from the users' perspective. However, it also has severe drawbacks: a browser is a big piece of software consuming a considerable amount of resources. Specifically, browsers tend to consume a lot of CPU and memory just to show a video on-screen. Considering that an important number of browsers will be needed in order to stress a media server, this is a huge use of resources. When these load testing strategies are to be considered in the cloud within a continuous integration pipeline, the cost of running so many browsers is also something to take into account.

Despite the drawbacks, this is the most common way of testing WebRTC applications, and the one in use in OpenVidu. There's a specific stress test performed frequently at the OpenVidu project: estimating the number of users supported by a single media node, which is a recurrent question in order to plan in advance the number of media nodes needed for a given load. To perform such a stress test, the master node is deployed on a machine in the cloud, and a single media node is started and attached to cluster.

Then, a Test Orchestrator is responsible for starting browsers and connecting them to the media node until this media node is not able to handle more connections. Usually this condition is detected by the Test Orchestrator as a browser request to join a session that does not receive any response. At this moment, the test finishes, and the Test Orchestrator reports the number of ses-

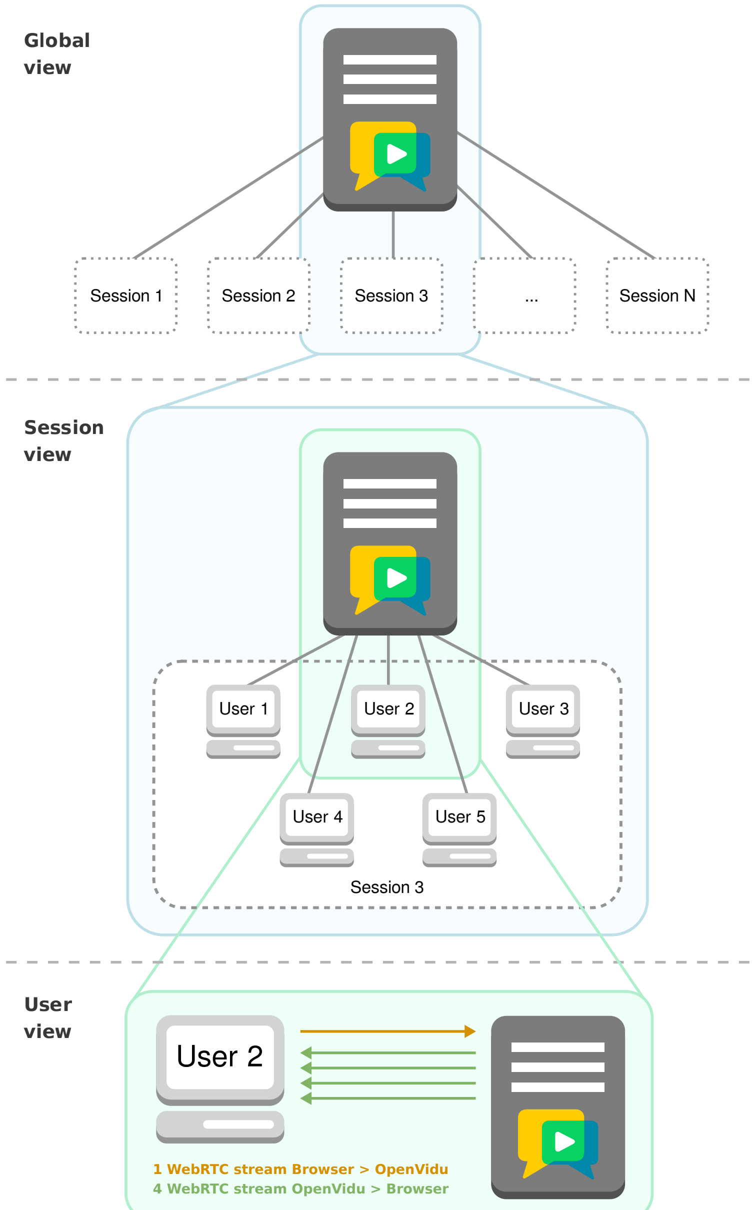

Figure 2: OpenVidu testing scenario for sessions with 5 users. Notice how each user sends a stream to the media node, and receives 4 streams from it (one per any other user in the session). The session is completely hosted in a single media node. The media node can host several of these sessions.

sions that were properly connected to the media server, i.e., those sessions for which all users successfully joined their respective sessions.

Usually, this stress test is performed on AWS, where machines running browsers can be requested on demand. As shown in Figure 2, the OpenVidu stress test starts EC2 instances (virtual machines) on demand in an AWS environment. Each instance hosts a single Chrome browser managed through Selenium. This strategy is very resource consuming, as many instances might be needed before it can be determined that the media server is not able to handle more connections. Even for media servers running on small instances, dozens of instances hosting browsers might be needed.

4. Load testing strategies for WebRTC applications

In this paper we propose new load testing techniques for WebRTC applications, and adaptations of existing ones, that optimize the use of resources, hence reducing costs for testing video conference applications. All the strategies depicted here were implemented and are available in the OpenVidu Loadtest Tool[7].

We made an effort within the OpenVidu Loadtest Tool to make strategies interchangeable, by sharing some pieces. Specifically, all the strategies run a user emulator service that exposes a REST API to the Test Orchestrator that can be used to join users to sessions (see Figure 3). The specificities of how users are joined depend on the different strategies as described below.

4.1. Selenium driven Browsers

The first strategy considered is the one described above of using web browsers to impersonate users in a video call. However, we can devise three slightly different ways in which this strategy can be applied that might result in different scalability, use of resources and costs: *browser*, *browser with recording* and *headless browser*.

[7]`https://github.com/OpenVidu/openvidu-loadtest`

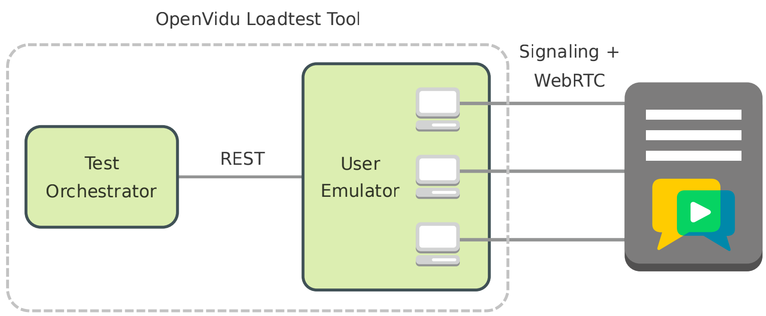

Figure 3: The emulator service presents a simple API to the Test Orchestrator and hides the specificities of the strategy being used.

In the most usual *browser* strategy we considered impersonating users using Chrome web browsers controlled through Selenium running in Docker containers as shown in Figure 4. Selenium is a library available in many programming languages that enables test code to programmatically control a browser (either running locally or remotely) through the browser's driver. When using this strategy, a new Chrome browser is started for each user joining the session. Browsers run a small test app, as shown in Figure 4 that uses the openvidu-browser library to join a specific session and sends a specific video and audio, instead of using the webcam and microphone, in order to have more control on the video being shared.

There's a second browser-based strategy, namely *browser with recording*, similar to the previous except that the open source ffmpeg[8] tool is used to record the Chrome browser window. This approach might require some more resources, as the ffmpeg tool is recording the window, encoding it in a suitable format and saving the video to disk. All while the browser is emitting and receiving video. Although in this paper we do not use the recorded videos in any way, we understand that in many cases these videos are used for acceptance testing. That's why we decided to consider recording the browser window.

Finally, there's still a third strategy involving browsers that consists on run-

[8] `https://www.ffmpeg.org/`

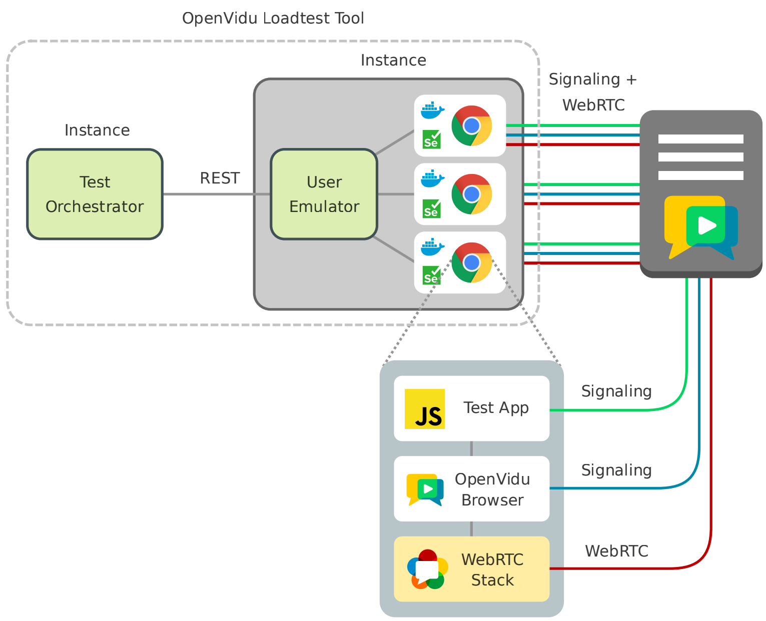

Figure 4: OpenVidu Loadtest Tool using browsers to impersonate users. Each browser is running in a Docker container and controlled through Selenium.

ning the browsers in headless mode. A *headless browser* is a standard browser without a UI. Everything within the browser works as usual except that the browser does not show the page on a graphical widget. We expect this headless mode to result in a reduction on the use of resources as page contents are not rendered on the screen.

In all these three strategies, each user is emulated through its own Chrome browser. The process is as follows: the Test Orchestrator sends a request to the user emulator service, specifying which session should the user join. The user emulator will then start a Docker container with a Chrome browser running a test application, configured to join the session specified. The test application leverages the openvidu-library to join the session and, once joined, a predefined video and audio available within the Docker container is sent to the media node.

4.2. Cost-effective testing solutions for WebRTC

In order to reduce the costs when load testing WebRTC services, we propose a new strategy and an adaptation of another one that was already described in [16]. Traditionally, browsers are used for WebRTC testing because they already implement the WebRTC APIs needed to establish and share media among peers. Therefore, it is relatively easy to build a small test application code to run in the browser that will join a specific session.

Our new proposal, *node-webrtc*, consists on using the WebRTC JavaScript library [9] that implements the WebRTC APIs. This library is able to impersonate many users from a single lightweight process (see Figure 5). To enable this library to send video and audio, we implemented a video and audio generator that generates and sends the media through the library.

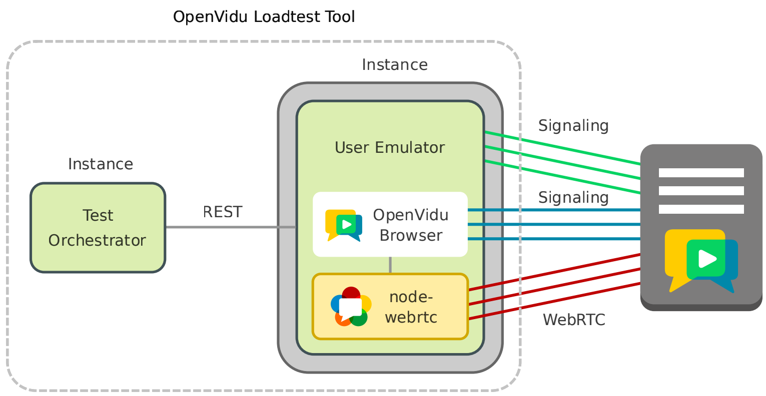

Figure 5: OpenVidu Loadtest Tool using the *node-webrtc* implementation to impersonate users.

This strategy is one of the main contributions of this work, and it is completely different to the previous ones. No browsers are used in this strategy. Instead, the *node-webrtc* library, that wraps the standard *libwebrtc* library for the Node platform, is able to impersonate users by directly connecting to a media node. The library implements parts of the WebRTC API, enabling the

[9] `https://github.com/node-webrtc/node-webrtc`

openvidu-browser library to use it as if it were the API of a real browser, thus doing the necessary configuration to send and receive video. With this strategy we get rid of browsers, thus we hypothesize that using this library should result in important savings in CPU and memory. Notice that *node-webrtc* library provides the standard WebRTC API found in browsers, so any WebRTC based library could be used instead of openvidu-browser with that strategy. Therefore, the results presented here can be considered valid regardless the media server used, as we are basically interested in reducing the testing costs implied by impersonating users with web browsers.

The second of our contributions is the *kms-webrtc* strategy, inspired in [8] and JAttack [16] that uses media servers instead of browsers to impersonate users. The main difference with previous strategies is that we implemented a library that wraps the specificities of the media server and provides the standard WebRTC API found in browsers, as in the *node-webrtc* strategy. Thus, the openvidu-browser library (or any other WebRTC signaling library) can be used unmodified. Other differences worth mentioning with respect to JAttack are the following: a) we used an unmodified standard Kurento Media Server, instead of a modified version of Janus media server (i.e., no modification of the media server was needed); b) video and audio generation is integrated in our proposal, reducing the complexity of the setup; and c) in order to use JAttack, a component called Controller has to be implemented to adapt the signaling to a specific media server (Janus). In the proposed OpenVidu Loadtest Tool the Test Orchestrator plays the role of the JAttack Controller. However, in our case the Test Orchestrator already implements typical load testing scenarios that can be selected with a high level configuration. The authors would have considered JAttack in our final experiment comparison, but this tool is closed source, and according to its creators it won't be open-sourced in a near future [10].

[10] https://groups.google.com/g/meetecho-janus/c/HwdfG82kZ_M/m/FVzq8dp1EAAJ

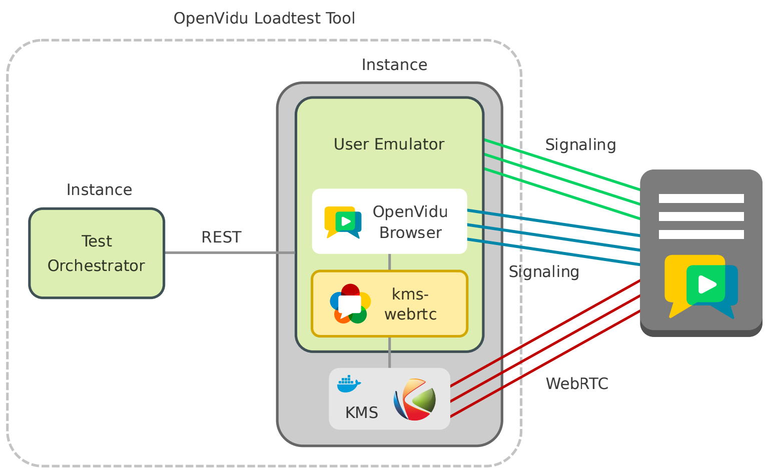

Figure 6: OpenVidu Loadtest Tool using the *kms-webrtc* strategy to impersonate users.

5. Methodology

Our comparison study has two parts: a preliminary experimentation and a final experimentation. In our preliminary experimentation we determined the limits of each testing strategy by answering RQ_{1a}, and RQ_{1b}. In our final experimentation, we studied the cost of the most promising strategy and compared it with a traditional strategy using browsers in a real stress test scenario, and used the results to answer RQ_2. Both experiments were performed in AWS, and consisted on: 1) an OpenVidu deployment (master node and media nodes), and 2) deployment of the necessary infrastructure for each strategy, as discussed in the rest of this section.

5.1. Preliminary experiment

First, we study how each of the proposed testing strategies perform in terms of: a) number of users that can be impersonated; and b) memory and cpu usage as we increase the number of users.

Given that the number of media streams exchanged grows up quadratically with the number of users in a session, as described in Section 3, we considered four different scenarios with different session sizes: 2, 3, 5, and 8 users per session. Each scenario has therefore a fixed session size (all sessions have the same size). Then, for each testing strategy, we run four different scenarios with a different session size each. Therefore, by performing the experiment with different session sizes we captured the overall behavior of the testing strategies and by analyzing the results we can answer RQ_{1a}. Notice that we had to run 20 tests in total, as we had to run four scenarios for each of the five testing strategies (*browser*, *browser with recording*, *headless browser*, *node-webrtc* and *kms-webrtc*).

As we are interested in the scalability of the different strategies, we considered a single AWS instance to impersonate all the users: the testing instance. The instance size chosen was a *t3.xlarge* with 4 vCPUs and 16Gb of memory, big enough to accommodate several users, at least with small session sizes. Preliminary results showed that this size was enough to ensure all strategies were able to hold at least one session in the biggest scenario (with 8 users per session). We used the same video for all the users emulated. The video had a resolution of 640x480, at 30 frames per second.

The OpenVidu platform (master and media nodes) was deployed on a sufficiently big instance so that the media nodes are never stressed (we are interested in estimating the number of users each testing strategy can handle in the given instance size). Specifically, we used c5.xlarge (4 vCPUs, 8 Gb) instances for both the master node and the media node. We carefully monitored the cpu and memory usage of media nodes to ensure they are never stressed.

All the instances used in the experiment were properly monitored, by running probes in each instance, and metrics were sent to and stored in an Elasticsearch[11] deployed in a different AWS instance. Data stored in this Elasticsearch was used for answering all RQs, and has been exported and made available in our reproduction package.

Each of the 20 tests to run consisted of a testing code that proceeded as shown in Algorithm 1.

[11] `https://www.elastic.co/elasticsearch/`

Algorithm 1 Preliminary experimentation testing procedure

```
procedure getUsersAndSessionsCreated(t, sessionSize)
    p ← 0                                         ▷ Number of users
    ns ← 0                                        ▷ Number of sessions
    stop ← false
    while !stop do
        sessionId ← newSession()
        i ← 0
        for i ← 1, sessionSize do
            stop ← addUser(t, sessionId)
            if !stop then
                p ← p + 1
            else
                break
            end if
            wait()
        end for
        if !stop then
            ns ← ns + 1
        end if
    end while
    return (p, ns)
end procedure
```

Where t is the testing strategy, $sessionSize$ is the number of users for each session (2, 3, 5, or 8), p is the total number of users already added (across all sessions) and ns is the number of complete sessions. The variables p and ns are increased during the test execution each time a new user is added and a new session is completed, respectively. A session is considered complete whenever all users have been added to the session (i.e., if the scenario consists on 5 users per sessions, all 5 users have been successfully added). The *newSession* method creates a new session in the OpenVidu platform and returns its id. Then, we resort to add users to this new session until all *sessionSize* users have been added. If at any time *addUser* fails to add a new user, the test stops, and reports the number of users added and the number of complete sessions successfully created. Notice that when the test stops there is a session yet to be completed. The maximum number of sessions completed ns for the strategy t will be used in the second part of our study.

Notice that after we successfully complete a session we wait for a reasonable time. This is to give some time to a new connection to stabilize. If we immediately request a new user to be added, the request might fail even when there are still available resources for the new user. This is due to the instance still working on adding the new user and the new streams to the other users in the same session. Therefore, to avoid stopping the test before the machine has been exhausted, we wait for some time, then we resume with the next session.

As was mentioned above, the *addUser* method works differently depending on the strategy t to use. For browser-based strategies, it uses Selenium to start a new browser for the new user within the AWS instance. This method returns a value indicating the stop condition when: a) Selenium is not able to start the browser; or b) the browser is not able to join the session (the request times out). For *node-webrtc*, the *addUser* method uses the *node-webrtc* API to instruct the library to add a new user. Similarly, for the *kms-webrtc* strategy, the *addUser* method uses the KMS API to instruct KMS to connect to a session. The error conditions for these two strategies are the same as for browser-based strategies.

In order to answer RQ_{1b}, during the experiment we collected CPU time and memory consumption metrics, using a metricbeat agent [12] that was run in the testing instance.

5.2. Final experiment

In our final experiment we performed a stress test against the OpenVidu platform using two different testing strategies: the most common strategy (*browser*) and the most scalable one according to our preliminary experiment (*kms-webrtc*). The OpenVidu deployment consisted on a master node (c5.large instance with 2 vCPUs and 4 GB of memory) and two media nodes (also c5.large instances).

The stress test added sessions until no more sessions could be started, i.e., until the platform became unresponsive. Usually, the number of sessions needed to make the system unresponsive are much more than the number of sessions

[12] `https://www.elastic.co/beats/metricbeat`

that a single testing instance can hold. Therefore, for this experiment, several testing instances were used. In our final experiment, we run the two strategies mentioned above on a representative scenario, and recorded the total number of sessions and users added, the running time of the test, and the number of instances used in total to impersonate those users. The topology of the final testing experiment is shown in Figure 7.

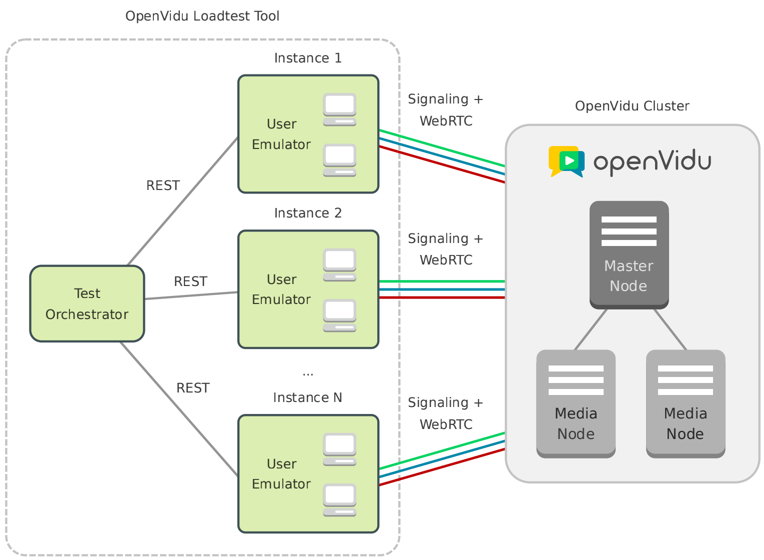

Figure 7: OpenVidu Loadtest Tool with several instances to perform stress testing against an OpenVidu cluster with two media nodes

Given we do not want to overload testing instances, we set the maximum number of sessions as the 70% of the value obtained from the preliminary experiment. We selected the scenario with $sessionSize = 5$. In that way using the *chrome* testing strategy we could reduce the maximum number of sessions from 2 to 1 and using the *kms-webrtc* we could reduce them from 12 to 8. We could not perform the comparison with $sessionSize = 8$ as the maximum number of sessions with strategy *browser* is 1.

In this experiment, the test proceeds as depicted in Algorithm 2. In this algorithm, t is the testing strategy, $sessionSize$ is the number of users per session for this test execution, $maxSessions$ is the maximum number of complete sessions to be started within each testing instance, and IPs is a list of IPs of available instances to be used to impersonate users. Notice that we started a sufficiently large number of instances in advance, and that the test does not start instances on demand (to save some time that otherwise would require starting each instance before being available). The algorithm then proceeds to initialize the number of users joined, sessions started and instances used to zero. Then, while users can be added it proceeds by choosing the first IP (line 8), and starting $maxSessions$ within that instance (lines 9–25), one session at a time (lines 10–24), and waiting until the session stabilizes (line 18), just like in our preliminary experiment. When all the $maxSessions$ have been started, we resort to the next instance (line 8). During the process we update the number of users (line 14), sessions completed (line 21) and instances used (line 7). The method $addUser$ is exactly the same as in Algorithm 1 and will return a code indicating a stopping condition under the same assumptions. The test returns the number of users that were successfully added, the number of complete sessions and the number of instances used. Externally, we also recorded the CPU wall clock time of the test execution.

With the outcome of this experiment we could calculate the cost of each testing scenario with the two testing strategies selected. As the cost of the master and media nodes is fixed, we considered exclusively the cost of the instances hosting the user emulators. We considered costs of the eu-west-1 AWS region where the experiments were run. All testing strategies made use of the same instance size (t3.xlarge), which had a cost of 0.1824 $ per hour, 0.00304 $ per minute at the time of the experiment. Let m_t be the number of test instances needed for strategy t until the test stops, $time$ be the time that it takes to complete the test (in minutes) and X be the cost of the instance (in \$/min), then the cost of strategy t can be calculated as $cost_t = time * X * m_t$.

Algorithm 2 Final experimentation testing procedure

```
procedure OPENVIDUSTRESSTEST(t, sessionSize. maxSessions, IPs)
    p ← 0                                          ▷ Number of users
    ns ← 0                                         ▷ Number of sessions
    ni ← 0                                         ▷ Number of instances used
    stop ← false
    while !stop do
        ni ← ni + 1
        IP ← IPs[ni]
        for j ← 1, maxSessions do
            sessionId ← newSession(IP)
            for k ← 1, sessionSize do
                stop ← addUser(IP, t, sessionId)
                if !stop then
                    p ← p + 1
                else
                    break
                end if
                wait()
            end for
            if !stop then
                ns ← ns + 1
            else
                break
            end if
        end for
    end while
    return (p, ns, ni)
end procedure
```

6. Experiment results

All the experiments were performed on AWS, using instances of different sizes as described below. We performed two different experiments. In our preliminary experimentation we tried to find the limits of each of the load testing strategies on a single AWS instance. In our final experimentation we tried to find the limits of the OpenVidu platform, using two of the strategies (the most common one and the best one from the preliminary experimentation) with the aim of compare the spending of each.

We always used the same instance size for the testing strategies (i.e., all strategies have the same number of CPUs and memory available), and the same video (i.e., all users are sending the same video), except the *node-webrtc* strategy that could not send a real video, and a new fake one was specifically generated for it.

6.1. Preliminary experiment

In the preliminary experiment we considered the five strategies described in Section 4, namely *browser*, *browser with recording*, *headless browser*, *node-webrtc* and *kms-webrtc*. We deployed the OpenVidu platform, with a single master node and a single media node. Both the master and the media nodes were deployed on a c5.xlarge AWS instance, with 4 vCPUs and 8 GB of memory. The strategies were all run in their own t3.xlarge instance with 4 vCPUs and 16 GB of memory.

Table 1 reports the session size (number of users per session), the testing strategy, the maximum number of completed sessions that each strategy could manage for each session size before starting to error when adding new users, and the maximum number of users that successfully joined their sessions. The number before the brackets denote the number of users considering only those that belong to a complete session, whereas the number in brackets denote the actual number of users, of which some might belong to a session not yet completed. For instance, for session size 2 the strategy *browser with recording* reported 10 users (corresponding to the 5 users in each of the 2 complete sessions), but 11 were able to join. The last one corresponds to a new, incomplete, session that only one user could join before the stop criteria was met.

As shown in the table, the number of complete sessions that a t3.xlarge instance can hold decreases with the session size in all the strategies. In general, using browsers we were able to generate less load than using our proposed strategies *node-webrtc* and *kms-webrtc*. When the session size goes beyond 3 users, that is, in the case for 5 and 8 users per session, the number of sessions supported by any of the browser-based strategies remain the same (2 and 1 respectively), indicating that there's little overhead in using the UI or recording

Session Size	**Testing strategies**	**Max complete Sessions**	**Max users**
2	browser	6	12 (12)
	browser w/ recording	5	10 (11)
	headless browser	7	14 (14)
	node-webrtc	18	36 (37)
	kms-webrtc	64	128 (129)
3	browser	4	12 (12)
	browser w/ recording	3	9 (10)
	headless browser	4	12 (14)
	node-webrtc	13	39 (40)
	kms-webrtc	26	78 (78)
5	browser	2	10 (12)
	browser w/ recording	2	10 (10)
	headless browser	2	10 (11)
	node-webrtc	5	25 (26)
	kms-webrtc	12	60 (62)
8	browser	1	8 (13)
	browser w/ recording	1	8 (10)
	headless browser	1	8 (11)
	node-webrtc	1	8 (15)
	kms-webrtc	2	16 (21)

Table 1: Results for the preliminary experiment

the browser. Therefore, when considering exclusively browser-based strategies, one could choose browsers with recordings, with the benefits of having a recording of each browser window.

The differences between the browser-based strategies and the other two strategies are huge, with *kms-webrtc* being the strategy that is able to generate more load (handle more complete sessions) in a single machine. This is even more clear in Figure 8 where the number of sessions supported by each strategy can be graphically compared. The *kms-webrtc* strategy systematically outperforms any other strategy no matter the session size. Notice how for a session size of 8, *kms-webrtc* successfully joined 21 users, more than 2 and a half sessions. For small session sizes *kms-webrtc* is able to handle 6 to 10 times more sessions than browser-based strategies. Our *node-webrtc* proposal also outperforms strategies based on browsers (except for session size 8 where they tie), handling 2 to 3 times more sessions than browsers.

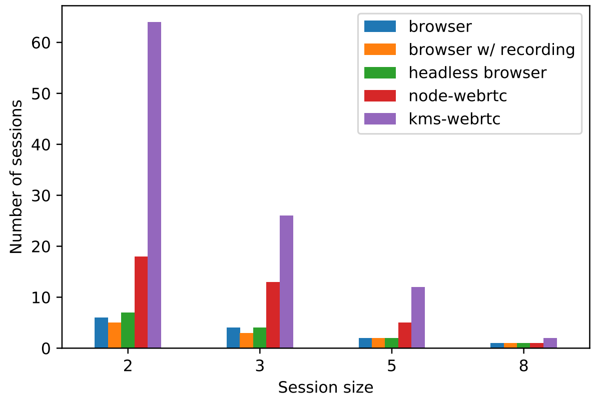

Figure 8: A comparison of the number of sessions supported by each strategy.

> **$\mathbf{RQ_{1a}}$: "What's the strategy that is able to emulate more sessions? Is this consistent with the session size?"** The two strategies proposed that avoid the use of browsers outperform any browser-based strategy. The strategy with better scalability (the highest number of sessions and users) is *kms-webrtc*. It is able to handle 64 sessions of 2 users, 26 sessions of 3 users, 12 sessions of 5 users and 2 sessions of 8 users, and it is consistent with the session size, being always above any other strategy. When compared to the most usual strategy of using browsers, *kms-webrtc* is able to handle at least twice the sessions of the browser-based strategies. For small session sizes (2 and 3 users), it is able to handle about an order of magnitude more sessions.

In addition to retrieve the number of sessions and users, each AWS instance was running a probe for monitoring CPU and memory usage. In Figure 9 we reported the usage of resources for each session size and strategy. We present the evolution of both metrics with the number of users.

Results show that *kms-webrtc* is again consistently below any other strategy in terms of CPU and memory usage. When considering session size 8, *node-webrtc* begins with a lower CPU and memory consumption, but as the number

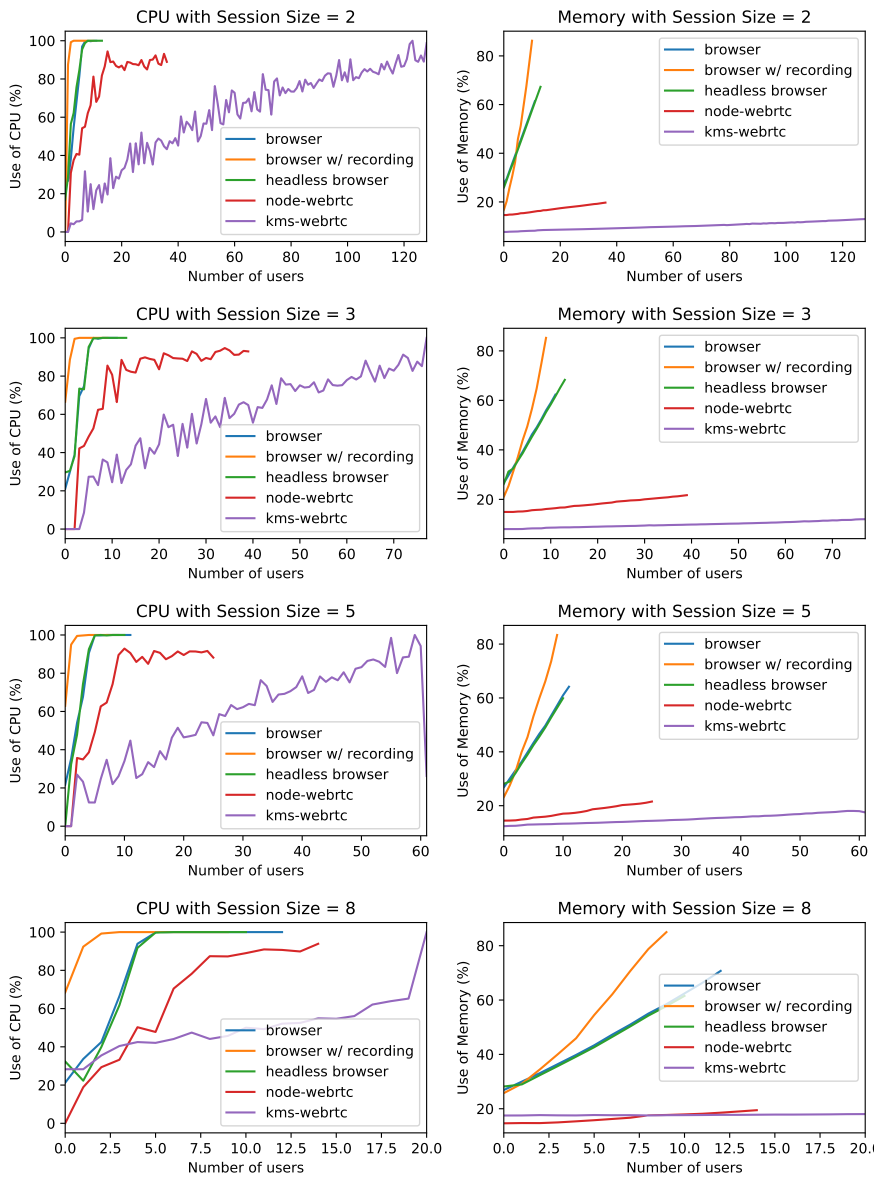

Figure 9: A comparison on the use of resources (CPU and memory) by session size for each of the strategies.

Strategy	Max complete sessions	Max users	Workers	Test time	Cost
browser	20	100 (104)	21	12 min	0.76608 $
kms-webrtc	24	120 (122)	3	7min	0.06384 $

Table 2: Results from our final experiment comparing costs of *browser* and *kms-webrtc* strategies.

of users increase, *kms-webrtc* stays below *node-webrtc*. Browser-based strategies quickly get to a situation where the CPU is at 100%, and their memory consumption grows quickly, hence the differences in Table 1.

$\mathbf{RQ_{1b}}$: “How do each of the strategies perform in terms of CPU and memory with respect to the number of users?” The two proposed strategies *node-webrtc* and *kms-webrtc* perform much better than any browser-based strategy in terms of memory. However, in terms of CPU, *kms-webrtc* performs much better than any of its competitors. When looking at the CPU, *node-webrtc* performs similarly to browser-based strategies. The steeper curves that strategies based on browsers expose in terms of memory are a clear indication that their scalability is poor, which means that not many browsers can be used within a single instance.

6.2. Final experiment

For our final experiment, we want to answer RQ_2. Therefore, we study the costs of a stress test using the usual browser-based strategy and the *kms-webrtc*, which is the strategy that performed best in our preliminary experiment. We used the same instance size for impersonating users as in the preliminary experiment: a t3.xlarge, but in this case the Test Orchestrator used several of these instances for emulating clients. For each instance, according to the 70% rule presented in Section 5.2, we used 1 session for *browser* and 8 sessions for *kms-webrtc*. The OpenVidu deployment consisted on one master node and two media nodes, all three nodes being c5.large instances (2 vCPUs and 4 GB of memory).

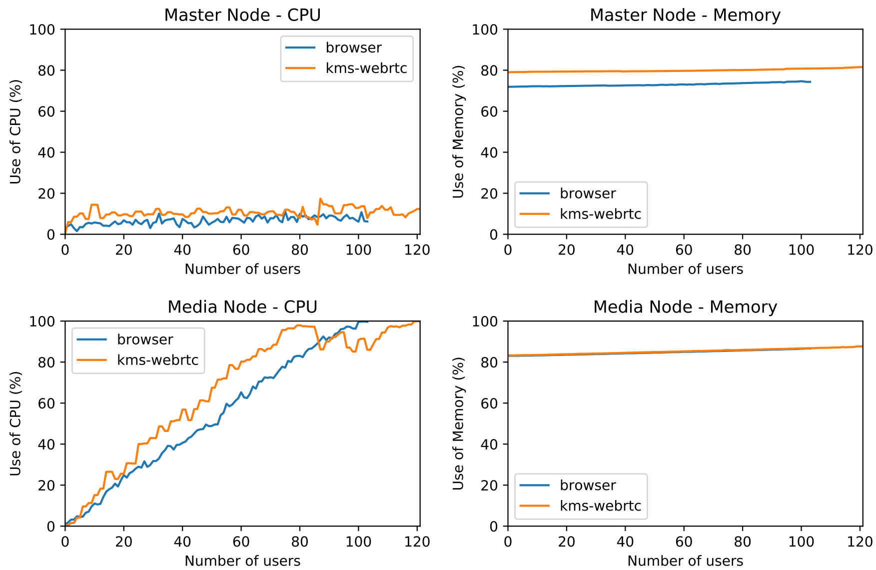

Figure 10: CPU and memory usage of the master node and the two media nodes (averaged).

Table 2 shows the results for the final experiment. For each strategy and session size the table reports the number of complete sessions before the OpenVidu platform started to fail, the number of users (the number between brackets is to be interpreted as in Table 1), the number of workers (number of testing instances that were used to host the sessions), the test time, and the cost. The cost is calculated as described in Section 5.2. The browser strategy needed 20 complete sessions to stress the OpenVidu platform. To host all those sessions 21 workers were required, running for a total time of 12 minutes. The total cost of the stress test with *browser* (rounded to the nearest hundredth) was 0.77\$. When using the *kms-webrtc* strategy, 24 complete sessions were needed to stress the OpenVidu platform. These 24 sessions were hosted in 3 instances, with a total expenditure of 0.06\$. This means a 7% of the spending of the test using browsers.

In order to ensure a fair comparison, we recorded the CPU and memory usage of the master and media nodes. Graphs in Figure 10 report these metrics, where data of the media nodes is averaged over the two nodes (hence a single graph for both media nodes). The CPU on the master node is below 20%, something expected as the master node is not handling the media, just forwarding users to one of the media nodes. However, the average CPU of the media nodes increases with the number of users. The behavior is similar in both strategies. The memory consumption is about 80% in the master node, without many variations. In the media nodes, it remains slightly above 70% and slowly increases with the number of users. As shown in the graphs, media nodes are CPU bound, that is, the number of users managed by a media node have a bigger impact on CPU than in memory.

> **$\mathbf{RQ_2}$: "Are there any cost savings for using our proposals with respect to using browsers when doing stress testing of a WebRTC platform?"** Definitely, there is a huge cost saving by using *kms-webrtc*. Expenditures can be cut off by 93%, due to using much fewer instances: 3 instances were needed by our proposal, 21 instances were needed by the usual strategy of using browsers.

6.3. Threats to validity

Our studies are subject to construct validity issues, mostly due to the load testing strategies chosen. The most common testing strategies use a browser to impersonate users, sometimes recording the browser window. In the state of the art we could not find any reference to the use of headless browsers, nevertheless, we considered this an interesting strategy. Authors believe the strategies described for load testing WebRTC services are the usual ones.

Our results are also subject to internal validity issues, due to how browsers adapt to environmental changes, mainly CPU available. In our preliminary experimentation, when the stop criteria is met, the test instance is reporting a CPU at 100%. This might cause the quality of the video to drop, and hence, the load over the media nodes decreases accordingly. To avoid this issue, in our final experiment we decided to use a value for the maximum number of sessions to run per testing instance equal to 70% of the maximum number of sessions reported for that strategy in our preliminary experiment. Additionally, there is also an issue with the video in the *node-webrtc* proposal, which is a synthetic one, as this proposal is not able to use a video stored locally as the video stream to be sent to the OpenVidu platform.

Finally, we also have external validity issues. We run our experiments in AWS instances (virtual machines) that possibly share the host machine with other instances (not owned by us). When the host is under a huge load, the situation might somehow impact the metrics we took during the test executions. The metrics we took, however, are consistent with what we expected, and we do not think that this issue materialized.

7. Conclusions

In this paper we have proposed cost-effective testing strategies for load testing WebRTC applications, and we have studied their scalability when compared with the usual strategies in the field, and their cost.

The main contributions of this paper are:

- A discussion of the most common browser-based strategies for load testing videoconference applications based on the WebRTC standard, and under which circumstances one strategy should be chosen over another.
- A proposal of a new strategy, *node-webrtc*, not based on browsers, that is more efficient in terms of use of resources, number of users emulated, than the browser-based ones.
- An improvement over a previous proposal [16] based on the usage of a media server to emulate users. Our strategy, *kms-webrtc*, includes improvements like adapting our strategy to the WebRTC APIs, or removing the need to implement ad-hoc testing code, among others. This strategy outperforms any other strategy in terms of number of users emulated.
- An open source load testing tool (the OpenVidu Loadtest Tool) that automates the process of generating load through any of these five strategies

to an OpenVidu deployment.

- Some results on the scalability of each strategy in terms of number of users, as well as CPU and memory usage.
- Some results on the monetary costs of two strategies (*browser* and *kms-webrtc*) when performing a stress test of a WebRTC platform. Results show 93% savings when using our proposal.

We tried to perform the studies in similar conditions for all the strategies. However, it might be interesting to research, for instance, how different methods behave under different conditions (like network issues), or how quality of experience degrades with load, both in the client (the browser may drop the quality of the video sent to avoid issues) and in the server (the OpenVidu platform might as well do some quality control to avoid congestion on peek loads).

Acknowledgements

A repository with the OpenVidu Loadtest Tool repository is available in GitHub[13]. Additionally, a reproduction package [22] is available in Zenodo with raw results and a copy of the OpenVidu Loadtest Tool repository.

This work has been supported by the Government of Spain through project "BugBirth" (RTI2018-101963-B-100), by the Regional Government of Madrid (CM) through project EDGEDATA-CM (P2018/TCS-4499) cofunded by FSE & FEDER, and by the European Commission under the H2020 project "MICADO" (GA-822717).

References

[1] Webrtc, `https://webrtc.org/` (2021 (accessed February 26, 2021)).

[2] Bigbluebutton, `https://bigbluebutton.org/` (2021 (accessed February 27, 2021)).

[13] `https://github.com/OpenVidu/openvidu-loadtest`

[3] Jitsi, https://meet.jit.si (2021 (accessed February 27, 2021)).

[4] Whereby, https://whereby.com/ (2021 (accessed March 22, 2021)).

[5] B. García, M. Gallego, F. Gortázar, E. Jiménez, Webrtc testing: State of the art., in: ICSOFT, 2017, pp. 363–371.

[6] A. Gouaillard, L. Roux, Real-time communication testing evolution with webrtc 1.0, in: 2017 Principles, Systems and Applications of IP Telecommunications (IPTComm), IEEE, 2017, pp. 1–8.

[7] A. Bertolino, A. Calabró, G. De Angelis, F. Gortázar, F. Lonetti, M. Maes, G. Tuñón, Quality-of-experience driven configuration of webrtc services through automated testing, in: 20th International Conference on Software Quality, Reliability and Security (QRS), IEEE, 2020, pp. 152–159.

[8] B. Garcia, F. Gortazar, L. Lopez-Fernandez, M. Gallego, M. Paris, Webrtc testing: challenges and practical solutions, IEEE Communications Standards Magazine 1 (2) (2017) 36–42.

[9] Openvidu, https://openvidu.io (2021 (accessed February 27, 2021)).

[10] A. Bertolino, Software testing research: Achievements, challenges, dreams, in: Future of Software Engineering (FOSE'07), IEEE, 2007, pp. 85–103.

[11] Y.-F. Li, P. K. Das, D. L. Dowe, Two decades of web application testing—a survey of recent advances, Information Systems 43 (2014) 20–54.

[12] G. A. Di Lucca, A. R. Fasolino, Testing web-based applications: The state of the art and future trends, Information and Software Technology 48 (12) (2006) 1172–1186.

[13] B. García, L. López-Fernández, M. Gallego, F. Gortázar, Testing framework for webrtc services, in: Proceedings of the 9th EAI International Conference on Mobile Multimedia Communications, 2016, pp. 40–47.

[14] Kurento media server, https://kurento.org (2021 (accessed February 27, 2021)).

[15] S. Taheri, L. A. Beni, A. V. Veidenbaum, A. Nicolau, R. Cammarota, J. Qiu, Q. Lu, M. R. Haghighat, Webrtcbench: a benchmark for performance assessment of webrtc implementations, in: 2015 13th IEEE Symposium on Embedded Systems For Real-time Multimedia (ESTIMedia), 2015, pp. 1–7. doi:10.1109/ESTIMedia.2015.7351769.

[16] A. Amirante, T. Castaldi, L. Miniero, S. Romano, Jattack: a webrtc load testing tool, in: 2016 Principles, Systems and Applications of IP Telecommunications (IPTComm), IEEE, 2016, pp. 1–6.

[17] A. Amirante, T. Castaldi, L. Miniero, S. P. Romano, Janus: a general purpose webrtc gateway, in: Proceedings of the Conference on Principles, Systems and Applications of IP Telecommunications, 2014, pp. 1–8.

[18] testrtc, https://testrtc.com/ (2021 (accessed March 17, 2021)).

[19] J. Gao, X. Bai, W.-T. Tsai, Cloud testing-issues, challenges, needs and practice, Software Engineering: An International Journal 1 (1) (2011) 9–23.

[20] A. Bertolino, G. D. Angelis, M. Gallego, B. García, F. Gortázar, F. Lonetti, E. Marchetti, A systematic review on cloud testing, ACM Comput. Surv. 52 (5) (Sep. 2019). doi:10.1145/3331447.

[21] E. André, N. Le Breton, A. Lemesle, L. Roux, A. Gouaillard, Comparative study of webrtc open source sfus for video conferencing, in: 2018 Principles, Systems and Applications of IP Telecommunications (IPTComm), IEEE, 2018, pp. 1–8.

[22] F. G. Bellas, M. G. Carrillo, M. M. Bermejo, I. Chicano, C. Santos, Dataset of paper "Cost-effective load testing of WebRTC applications" (Mar. 2021). doi:10.5281/zenodo.4661963.