

Wainhouse Evaluation- Enterprise Meetings | Video (Q4 2020)

Wainhouse quality evaluation of four market-leading enterprise meeting solutions – methodology and video results

Bill Haskins

In Q4 2020, Wainhouse evaluated four market-leading enterprise meeting solutions: Google Meet, Microsoft Teams, Cisco Webex Meetings, and Zoom Meetings. Our primary objective was to quantify and compare audio and video quality between each platform, using established and accepted industry standards to remove subjective user or tester influence.

We performed this evaluation in our tightly controlled lab environment following a repeatable process adapted from our existing evaluation methodology and documented guidance for the metrics we employed during testing.

This report summarizes the video component of this evaluation, including details on our methodology, the solutions we tested, and key findings.

Note: this evaluation was commissioned by Zoom. However, the findings provided in this report are unbiased and represent Wainhouse testing results and related perspectives on the topic. All these platforms are under constant improvement, and it should be recognized that what Wainhouse tested in Q4 2020, is likely to be different today. The reader should evaluate this, and all other evaluations against their own unique environment and enterprise requirements.

Wainhouse Quality Evaluation – Video

In Q4 2020, Wainhouse to evaluated four market-leading enterprise meeting solutions:

- Google Meet
- Microsoft Teams
- Cisco Webex Meetings
- Zoom Meetings

Our primary objective was to quantify and compare audio and video quality between each platform, using established and accepted industry standards to remove subjective user or tester influence. We performed this evaluation in our tightly controlled lab environment following a repeatable process adapted from our existing evaluation methodology and documented guidance for the metrics we employed during testing.

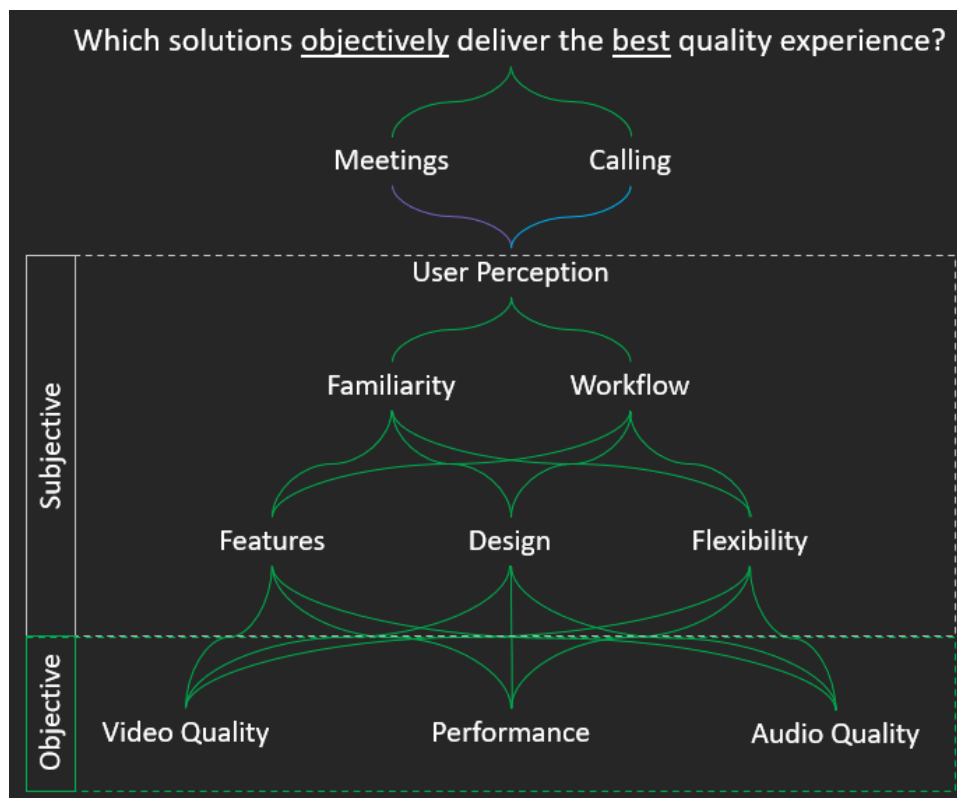
This brief summarizes the video component of this evaluation, including details on our methodology, the solutions we tested, and key findings.

Executive Summary

Evaluating and comparing quality across solutions has always been a tricky business. The word ‘Quality’, when associated with a meeting or calling solution, is generally interpreted as ‘*the user experience*’ – general, generic, broad, and personal. Many of the variables that make a user experience ‘great’ are, of course, subjective – experience, familiarity, design, and how a solution ‘fits’ within a user’s workflow. Each enterprise also has a unique set of criteria by which its compares meeting solutions: price, security, manageability, and so on.

But at the base of this conversation, we find an objective layer that serves as the foundation of the user experience – audio quality, video quality, and baseline performance. If a solution does not deliver at this layer, it is unlikely to make its way up the subjective ladder – users are unlikely to adopt low quality solutions, and the enterprise is unlikely to deploy them.

Enterprise Calling and Meetings - Quality Attributes



Source: Wainhouse Evaluation Lab, Q1 2021

The objective foundation can be quantified and compared using existing and accepted industry standards. With a focus on Audio and Video, we selected the ViSQOL metric for audio, and the VMAF metric for video – both are described in more detail within this report. Our evaluation included baseline, packet-loss, and latency tests, with each being repeated multiple times to validate results.

In Q4 2020, we applied this methodology to evaluate solutions from the market-leading enterprise cloud vendors: 8x8, Cisco, Google, Microsoft, RingCentral, and Zoom. We performed two separate-but-related evaluations, one focused on Meeting quality, and one focused on Calling quality. This brief outlines our methodology and three key results from our **Meeting evaluation** with a focus on **video quality**.

Evaluation Timeframe

This evaluation ran from September through November 2020. However, the bulk of this effort consisted of environmental configuration, pre-testing, and control validation. The final results summarized in this report are from tests conducted in **the first two weeks of November 2020**.

Evaluation Methodology

Wainhouse employs a repeatable evaluation methodology following a detailed test script that is applied to each solution. Core elements include:

Lab Environment

The WH lab consists of a set of control endpoints, distributed across multiple locations – West Chester, Ohio and Boulder, Colorado were the primary locations for this evaluation.

The lab employs a set of endpoints intended to reflect a common enterprise environment – for this evaluation, WH primarily used a relatively new (12-months-old) desktop, a relatively new laptop, and an aging desktop for quality- and performance-based tests. Additional endpoints were used when required (e.g., audio/video file capture), falling within a similar hardware / aging footprint as described in the table below.

Enterprise Calling and Meetings - Quality Attributes

| Lab Endpoint - Description | CPU | Memory | GPU | OS | WAN | Location |
|-----------------------------------|-----------------------------|---------------|-------------------|--------------------|--------------------------|--------------------|
| 1 yr-old Desktop - discrete GPU | AMD Ryzen 5 2400G | 16 GB | AMD Radeon RX 580 | Win10 Home (19041) | Cable 500MB down 20MB up | West Chester, Ohio |
| 1 yr-old Laptop - Integrated GPU | i7-1065G7 @1.3GHz (4 Cores) | 16 GB | Intel Iris Plus | Win10 Pro (19041) | Cable 1GB down 35MB up | Boulder, Colorado |
| 5 yr-old Desktop - discrete GPU | i7-5820K @3.3GHz (6 cores) | 32 GB | NVIDIA GTX 1070 | Win10 Pro (19041) | Cable 1GB down 35MB up | Boulder, Colorado |

Source: Wainhouse Evaluation Lab, Q4 2020

Environmental Controls

The lab environment is intended to reflect an average end-user’s environment – accessing cloud meeting services via the public internet, antivirus software enabled, etc. However, lab-specific controls were implemented to ensure a fair and consistent set of results across each solution, including:

Windows Resources: each endpoint was configured to reduce process and resource conflicts between tests: disabling shared storage / sync services and disabling indexing and antivirus for any folders sourcing or capturing video files. Note antivirus and indexing services were running through each test, but relevant folders were excluded from each. Endpoints were rebooted in between each test.

Time-of-day: all tests requiring public internet were conducted outside of business hours, after 5pm ET and on weekends.

Network: both send and receive / capture clients were in the same physical location, on the same network and subnet for each test. We did not enable QoS within this network for these tests.

Standards

We leveraged three video quality metrics (VMAF, SSIM, and PSNR) in this evaluation:

VMAF (Video Multimethod Assessment Fusion): designed by Netflix as a predictor of subjective video quality – it uses an AI-trained model to deliver a score aligned with an expected human ‘good-to-bad’ response. We selected VMAF as the primary video quality metric based on its trainable model, relevant controls, focus on video compression efficiency, and published research.

SSIM (Structural Similarity Index Measure): also a subjective predictor metric, but dated – it measures the structural similarity between reference and degraded video files, but lacks VMAF’s trained model.

PSNR (Peak Signal-to-Noise Ratio): designed to measure the quality of a reconstructed image as compared to an original reference image, looking for noise introduced during compression. This is the granddaddy video metric, but it is not well positioned to validate comparisons across different codecs and video technologies.

In all cases, VMAF is recognized as one of the best current predictors of human perception for video encoding technologies. While VMAF serves as our standard metric, we include SSIM and PSNR results as comparative metrics.

Process Detail

We established a 720p resolution as a standard baseline for all tests – this was the resolution ceiling for Google Meet and Webex Meetings; Zoom also provides control to limit resolution to 720p. Note that Microsoft Teams resolution ceiling is 1080p, and it does not provide a related limiting control.

We leveraged the following process for each video quality test:

1) Reference Files: We generated three 30-second reference files – a well-lit and dark office with a speaker talking and moving as expected in a ‘normal’ meeting environment, and a CGI-based motion file to push each platform’s motion handling capabilities. The files included a stopwatch and watermark to identify various test-flows (baseline, packet loss, and latency steps), and were rendered at 720p representing an average end-user webcam source.

2) Virtual Webcam: Each video was played at 720p and 30 frames-per-second using software with a native virtual webcam. We tested multiple solutions to ensure the highest quality across each meeting platform, using the same software for all video tests.

3) Endpoint Control: We cleaned (rebooted) each test PC and checked each meeting application for updates prior to every test cycle, making sure only the target test application was running at the time. All tests and sub-tests were completed using the same client version within each test category – e.g., video tests, audio VoIP tests, etc. We kept all solutions at their default quality settings except for Cisco Webex, where we disabled ‘Automatic Image Adjustment’ – we elected to disable this feature after pre-tests showed this setting to deliver materially lower scores.

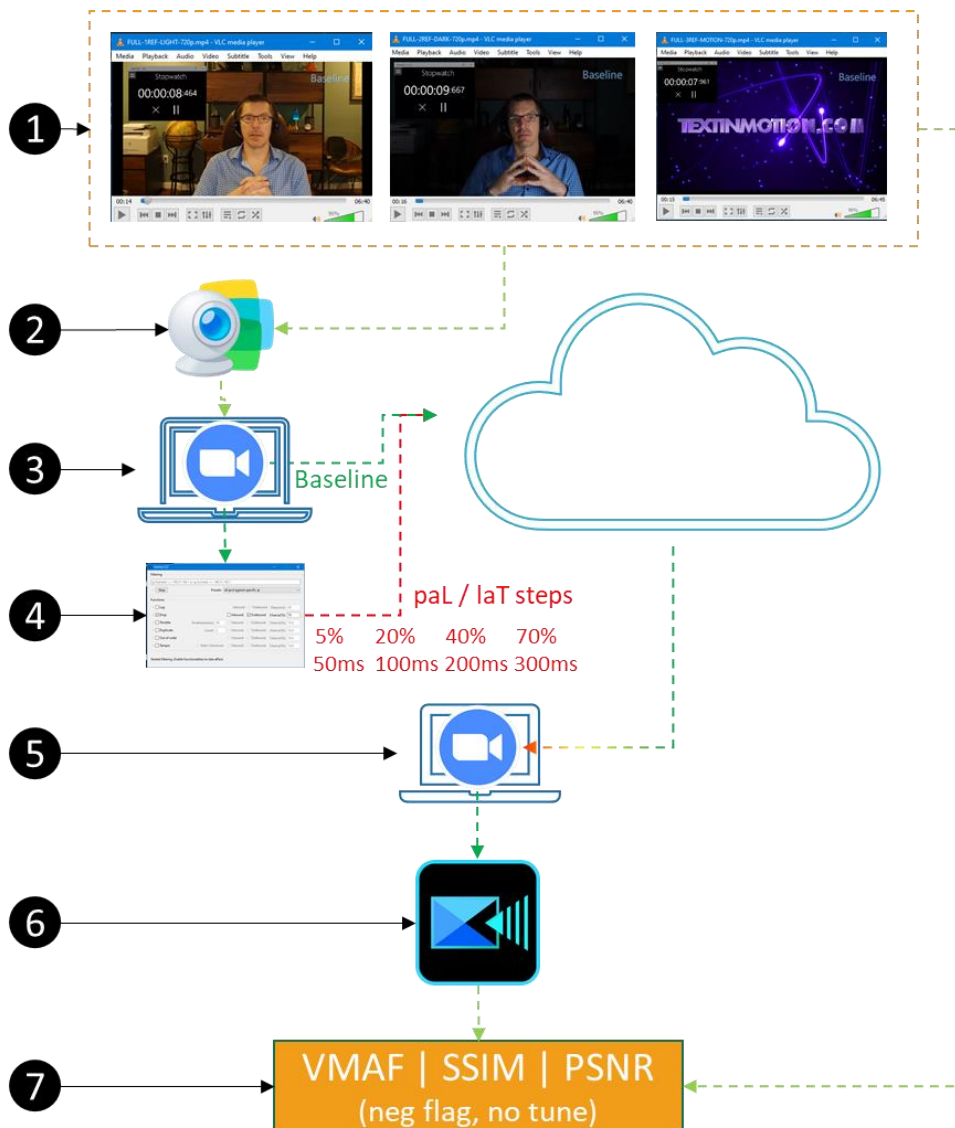
4) Network Conditions: We start each test under ‘normal’ network conditions (baseline), and then adjust network conditions across four packet-loss and latency steps – 3 files, 9 individual sub-tests, 27 total tests.

5) Capture Client: We play each video full-screen and use software to record the results – removing any additional overlay icons that may impact the scoring process.

6) Video Rendering: Each test file goes through a pre-scoring editing process – files are trimmed to the same ~30 second timeframe and ‘start-frames’ are aligned with the reference files. The results are rendered with a controlled setting across all solutions.

7) Scoring: Each ‘degraded’ test file is compared to the original ‘reference’ video using specialized scoring software – producing three individual scores for each test: VMAF, SSIM, and PSNR scores.

Video Quality – Test Methodology



Source: Wainhouse Evaluation Lab, Q4 2020

Evaluated Solutions

WH evaluated Google Meet, Microsoft Teams, Cisco Webex Meetings, and Zoom Meetings as detailed in the following table. The listed capabilities reflect each solution as tested at the time of this evaluation. During this evaluation, several solutions released updates that delivered a material improvement to a component of the test script (e.g., video quality, noise cancellation, etc.) When an update was determined to impact a given solution’s score, we re-tested all solutions within the relevant section of our test script for consistency.

In addition, we chose to leave as many settings at their default values within each solution when possible – anticipating that the average user does not actively configure or optimize their video settings. In some cases, however, we did make vendor- or service-specific changes – primarily if a solution-specific default setting delivered a material negative impact to its score. Additional solution details follow:

Google Meet: Evaluated on Chrome v. 86.0.x / 64 bit (no native client at the time of testing). Hardware Acceleration was enabled within each Chrome browser.

Webex Meetings: We evaluated the 40.10 client. We disabled Webex Meeting’s ‘auto-image-adjustment’ setting after pre-tests showed it delivered lower VMAF scores.

Zoom Meetings: We evaluated v. 5.3.2 and Zoom’s 720p experience as our primary test focus within this evaluation. We also conducted a series of secondary Zoom 1080p video quality tests, primarily to compare against Microsoft Teams scoring (the only other 1080p-capable solution in this evaluation). However, unless noted otherwise, Zoom’s video results refer to its 720p experience in this report.

Microsoft Teams: Evaluated v. 1.3.00.x (64-bit), default (1080p) settings.

Note WH evaluated solutions across a range of product tiers. The mix of product tiers was primarily a function of existing in-place Wainhouse tenants and purchasing / procurement requirements. In each case, Wainhouse evaluated each procured solution’s feature-set to ensure it aligned with the core Quality-focused requirements for this evaluation – primarily support for HD audio, video, and PSTN calling. Each solution was marketed and positioned as supporting the core required feature set, and Wainhouse therefore expects no change in related quality scoring based on the selected service tiers.

Meeting Evaluation – Service Description

| | Google Meet | Microsoft Teams | Webex Meetings | Zoom Meetings |
|----------------------|----------------------------------|-------------------------|------------------|---------------|
| Evaluated Plan | Google Workspace Enterprise Plus | Microsoft Office 365 E5 | Cisco Webex Plus | Zoom Pro |
| Evaluated Solution | Chrome v. 86.0.4240.111 | v.1.3.00.28779 (64-bit) | v. 40.10.6.11 | v.5.3.2 |
| Max Video Resolution | 720p | 1080p | 720p | 1080p |
| Video Codecs | VP9 | H.264 HW | H.264 AVC | H.264 AVC |

Note: capabilities reflect the plans and services used at the time of this evaluation.

Source: Wainhouse Evaluation Lab, Q4 2020

Summary Results

Baseline Scores

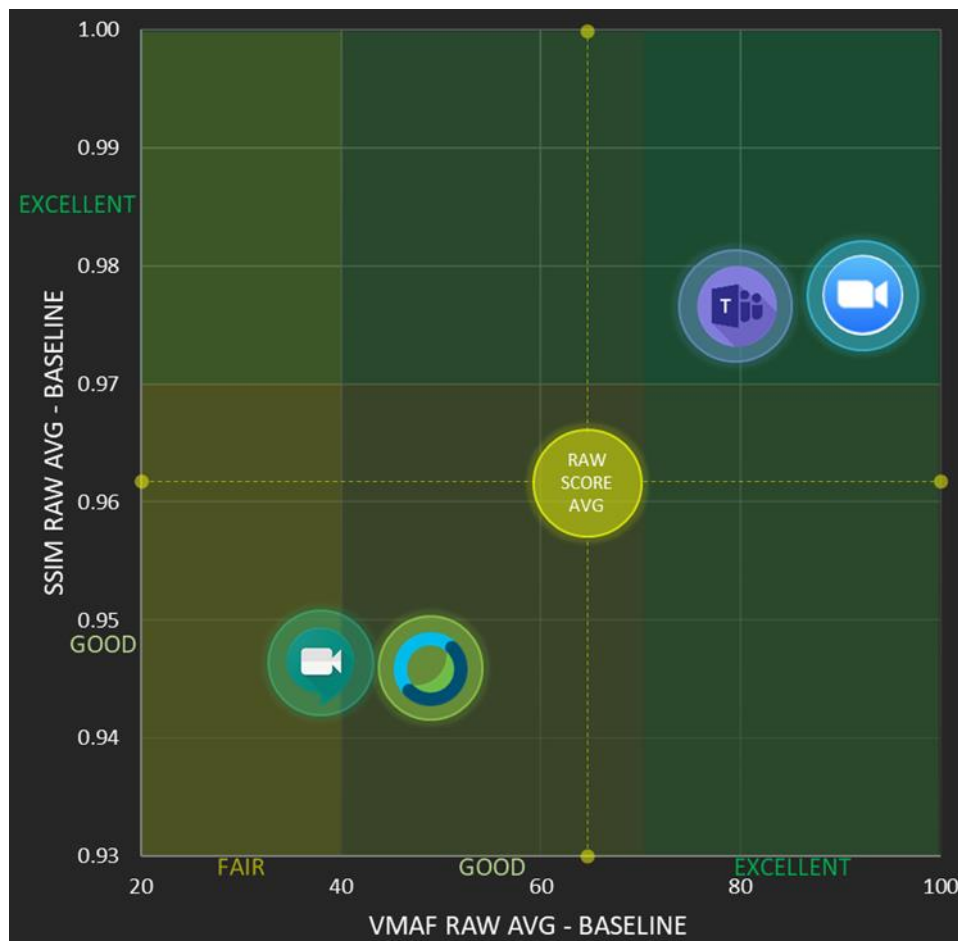
Description: This is a matrix chart that serves as our primary format to visualize quality-related results, comparing three related results within the same segment or test category. Here we are comparing each solution’s average evaluation score across all *Baseline Tests* for each Video Quality metric – with VMAF results on the X-axis, SSIM on the Y-axis, and PSNR results on the Z-axis. Rankings are based on average Baseline scores for each solution. Again, these are *Baseline* results only (no packet-loss or latency), averaged across Light, Dark, and Motion reference files. The X-axis represents a Primary attribute (VMAF = the primary metric for Video Quality in this evaluation), Y

represents a Secondary attribute (SSIM as our secondary metric), and Z a logical comparison attribute (comparing PSNR scores here).

Note the color pattern within this chart – these are a visual representation of the established quality range for each metric. These ranges follow a 1 (bad) to 5 (excellent) Absolute Category Ranking scale for each metric. Each metric includes a standard quality range – however, Wainhouse adapted and extended each range based on complexities and limitations that may have reduced or otherwise negatively impacted scores (e.g., virtual webcams, video capture, editing, and rendering processes). That said, all solutions were scored within the same controlled environment – we believe the scores provide a relevant view of each platform’s relative relationship.

Analysis: We see that Zoom delivered the highest baseline VMAF results, measurably higher than second-place Microsoft Teams in this evaluation. However, both solutions delivered a nearly identical SSIM score. (You will have to go to the fourth decimal place to find Zoom up by a .0009th of a point.)

Google and Cisco are similarly grouped in terms of VMAF and SSIM scores – although a) both scored relatively low here, and b) we see Google sliding into the ‘Fair’ range with its sub-40 VMAF score.



Source: Wainhouse Evaluation Lab, Q4 2020

Metric Comparison – Baseline Scores

| | Zoom | Microsoft Teams | Webex Meetings | Google Meet | AVERAGE |
|------|-------|-----------------|----------------|-------------|--------------|
| VMAF | 92.24 | 79.54 | 49.02 | 38.01 | 64.70 |
| SSIM | 0.98 | 0.98 | 0.95 | 0.95 | 0.96 |
| PSNR | 30.20 | 31.61 | 26.57 | 27.26 | 28.91 |

Source: Wainhouse Evaluation Lab, Q4 2020

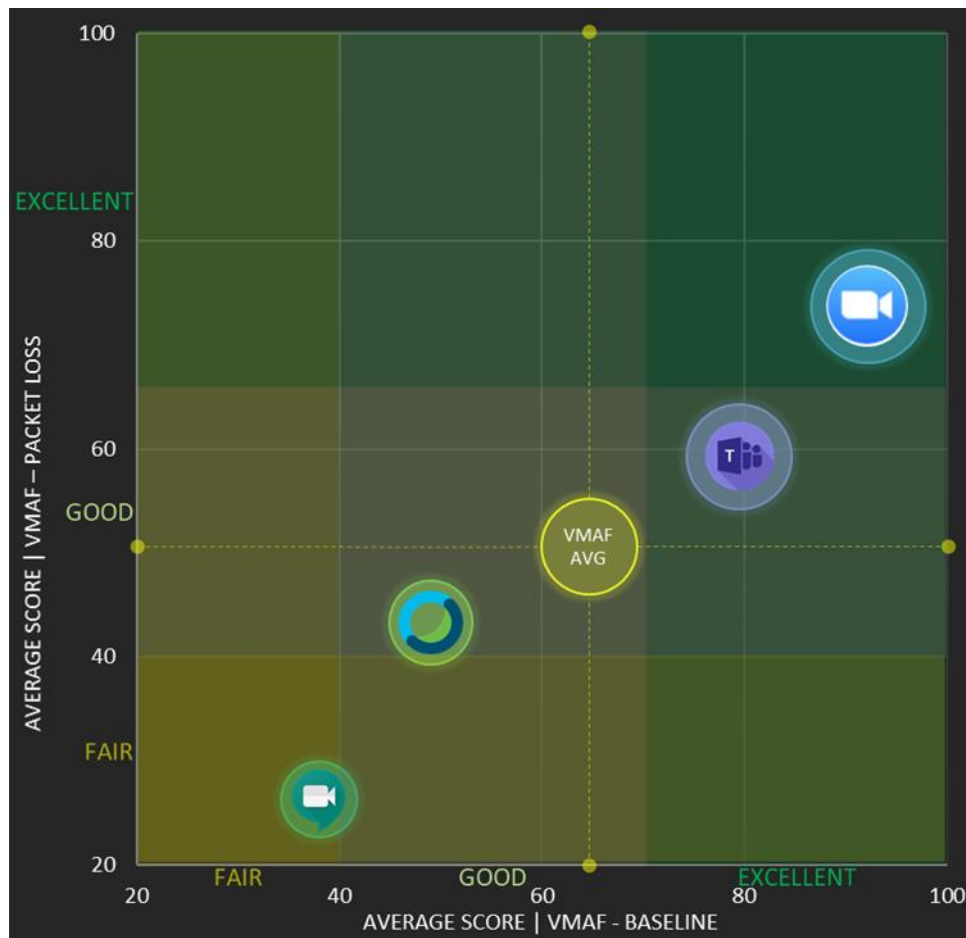
Network Degradation

Description: This chart plots average VMAF scores for all video quality tests (Light, Dark, Motion) by component, with Baseline results on the X-axis, Packet Loss on the Y-axis, and Latency on the Z-axis. Rankings are based on average VMAF scores across Baseline, Packet Loss, and Latency tests for each solution.

Note the X-axis here is the same as in the previous chart – plotting VMAF scores for baseline tests. So, we find each solution in the same horizontal (left-to-right) position. However, we have plotted average packet-loss scores on the Y-axis – these results average all packet-loss tests: 5%, 20%, 40%, and 70% steps. Similarly, each bubble size (Z-axis) represents latency results, across 50ms, 100ms, 200ms, and 300ms steps.

Analysis: The vertical position for each solution shows how well each scored while we dropped packets. Again, we see Zoom maintaining a material lead over the pack during packet-loss testing, followed by Microsoft and Cisco (both remain in ‘Good’ territory), with Google dropping down into ‘Fair’ territory during packet-loss testing.

VMAF Results – Baseline vs. Network Degradation



Source: Wainhouse Evaluation Lab, Q4 2020

VMAF Results – Baseline vs. Network Degradation

| | ZOOM | Microsoft Teams | Webex Meetings | Google Meet | AVERAGE |
|--------------------|-------|-----------------|----------------|-------------|--------------|
| Baseline | 92.24 | 79.54 | 49.02 | 38.01 | 64.70 |
| Packet Loss | | | | | |
| 5% | 89.67 | 73.31 | 46.31 | 33.87 | 60.79 |
| 20% | 84.41 | 67.25 | 46.67 | 33.44 | 57.94 |
| 40% | 73.23 | 50.37 | 44.72 | 17.46 | 46.44 |
| 70% | 47.52 | 46.20 | 35.53 | 20.26 | 37.38 |
| Latency | | | | | |
| 50ms | 84.49 | 74.88 | 45.56 | 37.87 | 60.70 |
| 100ms | 89.53 | 72.20 | 45.92 | 36.76 | 61.10 |
| 200ms | 78.34 | 68.47 | 44.27 | 34.99 | 56.52 |
| 300ms | 75.94 | 64.67 | 41.58 | 35.86 | 54.51 |

Source: Wainhouse Evaluation Lab, Q4 2020

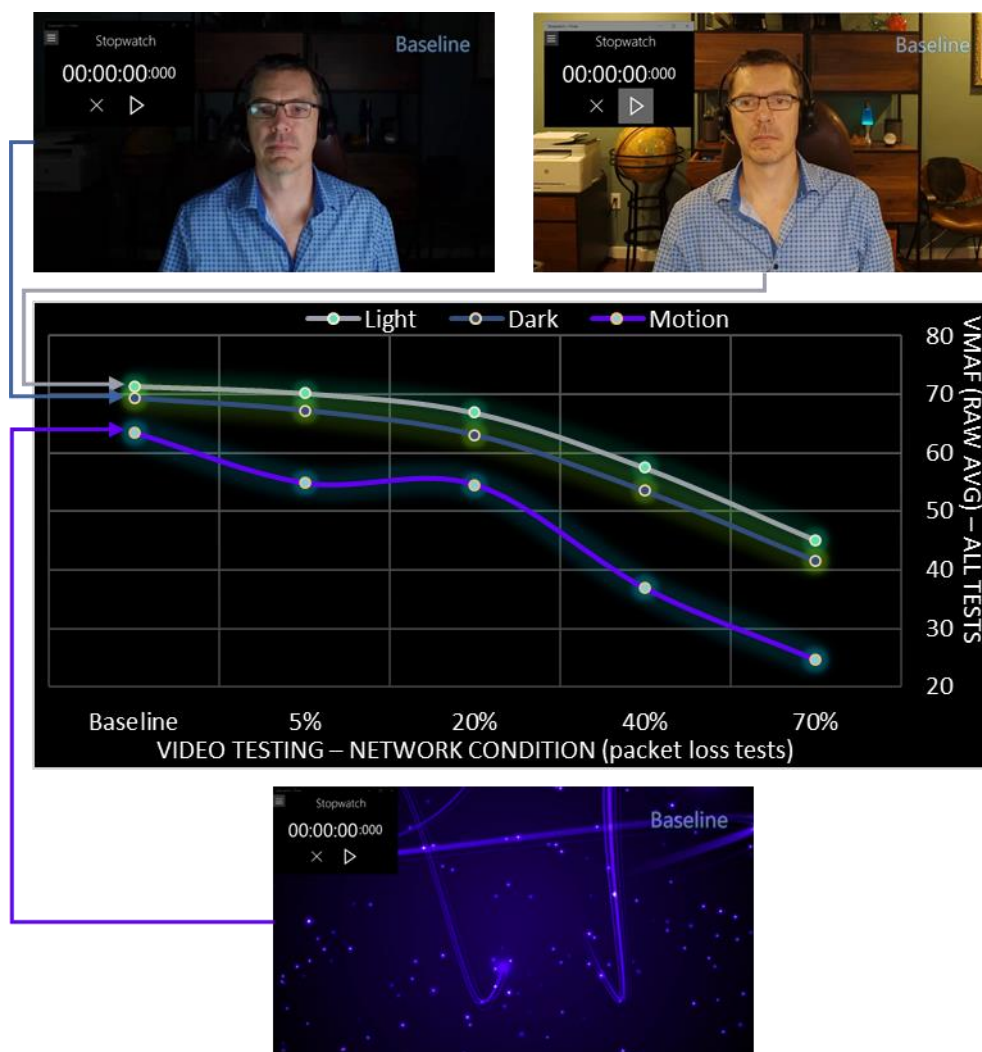
Motion Handling – Average Results

Description: We used three primary reference files for this video evaluation – testing quality in a light and dark environment, and with constant motion. This graphic shows an example of each reference file, with the average Baseline and Latency results across all solutions.

Analysis: These results show a consistent pattern between light and dark reference files – dark video generates slightly lower quality scores, and both have a consistently lower scores as more packets are dropped. Note the gap in scores increases slightly at 5% packet loss but stays fairly even from 20% to 70% loss. We also found it interesting that average scores stay relatively close between Baseline and 20% packet loss, then drop from 40% to 70% loss-rates – we expect loss above 10% is abnormal for the average video call, and it's logical for vendors to focus accordingly on maintaining quality in this <20% range.

We should also acknowledge that our Motion reference video does not reflect a 'normal' meeting experience. That said, we used this file as a way for us – and the video scoring engine – to understand how each solution handles extreme motion. That said, we see a larger delta between motion and light/dark results – no surprise, considering each solution needs to refresh more pixels and process more data during the motion test. Note the larger dip between Baseline and 5% loss, followed by a flat-line to 20% loss, then steep drop to 40% and 70% loss-rates; overall, the motion tests generated a larger variance between results, a variance that was amplified as we dropped packets.

VMAF Average Results – File Type Comparison



Source: Wainhouse Evaluation Lab, Q4 2020

Motion Handling – Solution Results

Description: This ‘Teardrop’ Chart compares average VMAF results across each reference video type – Light, Dark, and Motion. Average scores for each solution are plotted on the Y-axis (vertical layout). Rankings are based on average combined scores across all file types for each solution.

Analysis: Comparing average VMAF results by file type highlights a consistent trend: scores decline an average -9% between Light & Dark files, and a steeper -23% between Dark and Motion files.

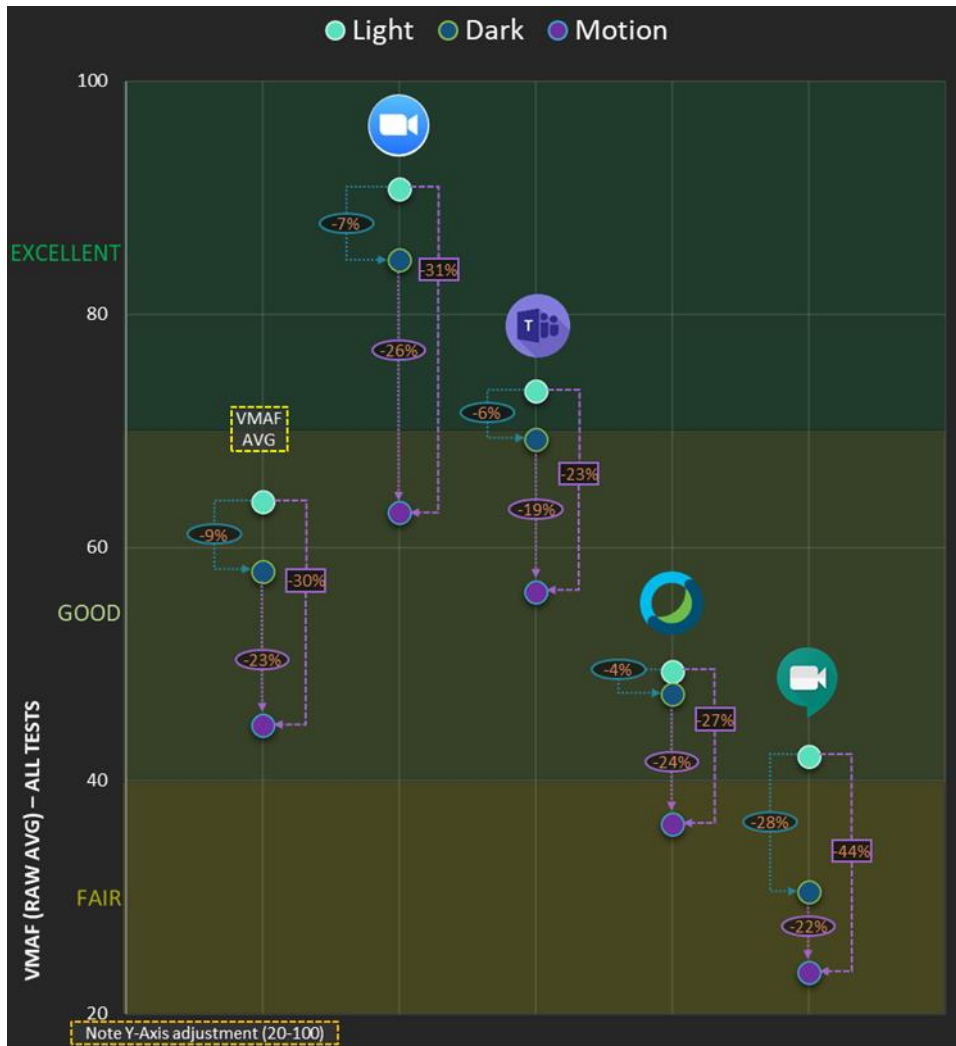
On the right side of this chart, you will see the Google bucking this trend, showing a steeper decline between Light and Dark than Dark and Motion – with the deepest 44% decline between Light and Motion scores. Compare this to Microsoft, a platform delivering the smallest decline in VMAF scores between file types. And while Cisco had low baseline scores, it also delivered the closest Light and Dark scores with a 4% decline between the two.

Microsoft delivered the smallest delta between file types – note the 6% decline between light and dark files, and the overall 23% decline between light and motion files. Again, this dataset includes

impact across all network degradation tests – we expect any Motion Sickness was primarily attributed to aggressive Packet Loss tests.

Zoom’s Light vs Dark scores were similar to Microsoft’s, but the Motion file pushed its overall VMAF score lower. Again, this dataset includes impact across all network degradation tests – we expect any Motion Sickness was primarily attributed to aggressive Packet Loss tests.

VMAF Results – File Type Comparison



Source: Wainhouse Evaluation Lab, Q4 2020

VMAF Results – File Type Comparison

| | ZOOM | Microsoft Teams | Webex Meetings | Google Meet | AVERAGE |
|--------|-------|-----------------|----------------|-------------|--------------|
| Light | 90.77 | 73.45 | 49.41 | 42.12 | 63.94 |
| Dark | 84.71 | 69.29 | 47.48 | 30.45 | 57.98 |
| Motion | 62.98 | 56.23 | 36.30 | 23.61 | 44.78 |

Source: Wainhouse Evaluation Lab, Q4 2020

Summary

Our goal in this brief is to provide transparency into our process and methodologies, and a few examples of the output our approach delivers. Our objective is to maintain objectivity and to continually improve our process – we welcome related questions and feedback on this note.

Keep in mind this evaluation was conducted in a controlled lab environment – we think the results are useful from a comparative point of view, but individual experiences will vary based on each unique environment. Also note, this evaluation was conducted at a point in time, and each vendor has been busy updating, iterating, and enhancing their platforms. Wainhouse will continue this series over time, keeping tabs on each solution as they evolve, retesting, re-evaluating, and reporting our findings as we move forward.

About Us

About the Authors

Bill Haskins is a Senior Analyst at Wainhouse Research with a strategic focus on unified communications products and services. Bill has over 15 years of experience supporting, delivering, and designing converged Collaboration services in a global communications environment. He has authored multiple white papers and articles detailing the keys to a successful UCC implementation and delivered various UCC presentations, highlighting his experience integrating Collaboration solutions into business process and enterprise applications.

Bryan L Hellard is a Researcher at Wainhouse where his primary focus is product evaluation and testing. He has 20 years of experience in the industry across several roles, including product engineering and management, R&D, and end user consulting. Prior to Wainhouse Research, he was President of True View Video where he developed video conferencing related products and consulted with end users on best practices for collaboration. Bryan has also been a consultant to video collaboration vendors providing product design services. He lives in the Cincinnati, Ohio area.

About Wainhouse Research

Wainhouse Research is an independent analyst firm that focuses on critical issues in the unified communications and collaboration market. The company provides 6 different vendor subscriptions covering unified communications, enterprise video, meeting room collaboration, personal & web-based collaboration, and audio conferencing, as well as a single all-inclusive subscription for enterprise users. The company acts as a trusted advisor providing strategic advice and direction for both the UC&C industry and its enterprise users. For further details contact sales@wainhouse.com or see <http://www.wainhouse.com>.

Notices

Copyright Notice

Copyright ©2021 by Wainhouse Research, LLC. All rights, including that of translation into other languages are specifically reserved. No part of this publication may be reproduced in any form, stored in a retrieval system, or transmitted by any method or means, electrical, mechanical, photographic, or otherwise, without the express written permission of Wainhouse Research, 34 Duck Hill Terrace, Duxbury, MA 02332 (Tel 781.312.6015) www.wainhouse.com. This publication is protected by United States copyright laws and international treaties. Unauthorized distribution or reproduction of this publication, or any portion of it, may result in severe civil and criminal penalties, and will be prosecuted to the maximum extent necessary to protect the rights of the publisher.

Limitation of Liability Notice

The information contained herein has been obtained from sources believed to be reliable. Wainhouse Research shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve its intended results. The opinions expressed herein are subject to change without notice.