The Impact of GOP Size on Video Quality

Jan Ozer
Streaming Learning Center
jan.ozer@streaminglearningcenter.com
November, 2023

Agenda

- Overview
- How I tested
- The command strings
- Impact of GOP Size on Quality - H.264
- Impact of GOP Size on Quality - HEVC
- Summary and conclusion

STREAMING
LEARNING CENTER

Online Course for New Streaming Professionals

Streaming Media 101:
Technical Onboarding for Streaming Media Professionals

Learn the technologies, techniques, and skills to succeed in a streaming media-related role, whether it's producing and distributing streaming video or creating the tools and services necessary to do so.

https://bit.ly/StreamingMedia101

What You Will Learn

In about 11 hours, this online course will teach you the terms, technologies, best practices, and skills needed to excel in a technical role in the streaming media industry. You will learn:

- How to encode and deploy streaming video using the H.264, HEVC, VP9, and AV1 codecs
- How to encode for single file and adaptive bitrate encoding and packaging for HLS, DASH, and CMAF
- About digital rights management (DRM) and distribution issues like choosing a CDN and how to measure and ensure Quality of Service and Quality of Experience
- Critical production-level decisions, like whether to encode on-premise or in the cloud, how to choose a per-title encoding technology and cloud encoder, and how to compute the breakeven on deploying an advanced codec like HEVC or AV1

You will learn to:

- Analyze files with MediaInfo, Bitrate Viewer, Apple's AVQT, and the Moscow State University Video Quality Measurement Tool
- Encode in FFmpeg and Handbrake
- Produce mezzanine files for upload to a streaming service
- Connect to YouTube Live and Facebook Live
- Embed a live or on-demand video into a web page

Overview

GOP Size	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
Overall HEVC	91.44	94.04	95.06	95.34	95.46	95.52	95.66	95.73
Max delta - HEVC	4.29	1.68	0.66	0.39	0.26	0.21	0.06	0.00
Overall H.264	88.73	92.42	93.92	94.32	94.49	94.61	94.79	94.86
Max delta - H264	6.13	2.43	0.94	0.54	0.37	0.25	0.07	0.00

One of the most fundamental encoding decisions is GOP size, or the frequency of I-frames in our encoded files. I-frames start a “group of pictures” comprised of I-, B-, and P-frames.

Most of the time our GOP size, or I-frame interval, is dictated by adaptive bitrate considerations like choosing a GOP size that divides evenly into your segment size. But even then, you have multiple options. In addition, what GOP size should you use when encoding a single file for disk-based playback or progressive download?

Some people believe that since I-frames are the highest quality frame, more I-frames are better. However, as the data shows above, it’s just the opposite.

That is, since I-frames are the *least* efficient, more I-frames *degrades* quality. As you would expect, the quality delta is most severe at very small GOP sizes, and if you’ve been using a GOP size of one second it’s time for a rethink. At the other end of the spectrum, the difference between ten and twenty seconds is irrelevant.

I measured the impact with H.264 and HEVC, and the delta was more significant with H.264.

I tested four video categories; animation, entertainment, sports, and office, and the quality delta at low GOP sizes was greatest for office content like talking heads, screencams, and PowerPoint tutorials.

STREAMING LEARNING CENTER

How I Tested

Whenever you measure the impact of configuration options on video quality, you should test within the most relevant quality range for actual publishers. For 1080p 24/25/30p videos like those tested here, the target VMAF score should be between 93-95. When testing a range of files, this means that you need to customize the bitrate for each file, which is time consuming, but unavoidable.

Here's the procedure I followed.

1. Choose a range of test files in multiple genres, which included animation, general entertainment, sports, and office. Files ranged in duration from about two to four minutes.

2. Encode at CRF 27 with x264 and x265; measure bitrate and VMAF. CRF 27 should yield a VMAF score in the target range, and you can adjust the bitrate accordingly during the actual test encodes depending upon the VMAF score.

3. Encode using two-pass 200% constrained VBR. Obviously, this is a VOD use case, though I'm guessing that live results wouldn't be that different.

Relevant configuration options you'll see on the following page include:

- Veryslow preset for both x264 and x265. See here.
- 200% constrained VBR. See here.
- Closed GOPs. See here.

I computed VMAF using the harmonic mean score using the Moscow State University Video Quality Measurement Tool. I did not track low-frame scores in this instance because I felt that the harmonic mean scores were sufficient to tell the tale.

STREAMING LEARNING CENTER

Command Strings

H.264

```
ffmpeg -y -i Freedom.mp4 -c:v libx264  -preset veryslow -force_key_frames expr:gte(t,n_forced*.5)
-threads 8 -an -b:v 3600K -maxrate 7200K -bufsize 7200K  -pass 1 -f mp4 NUL & \

ffmpeg -y -i Freedom.mp4 -c:v libx264  -preset veryslow -force_key_frames expr:gte(t,n_forced*10)
-threads 8 -an -b:v 3600K -maxrate 7200K -bufsize 7200K -pass 2 Freedom_264_GOP7_10.mp4
```

HEVC

```
ffmpeg -y -i freedom.mp4 -c:v libx265  -preset veryslow -x265-params
keyint=300:min-keyint=300:scenecut=0:bitrate=2500:vbv-maxrate=5000:vbv-bufsize=5000:frame-threads=8:
open-gop=0:pass=1 -f mp4 NUL & \

ffmpeg -y -i freedom.mp4 -c:v libx265  -preset veryslow -x265-params
keyint=300:min-keyint=300:scenecut=0:bitrate=2500:vbv-maxrate=5000:vbv-bufsize=5000:frame-threads=8:
open-gop=0:pass=2 freedom_HEVC_GOP7_10.mp4
```

STREAMING LEARNING CENTER

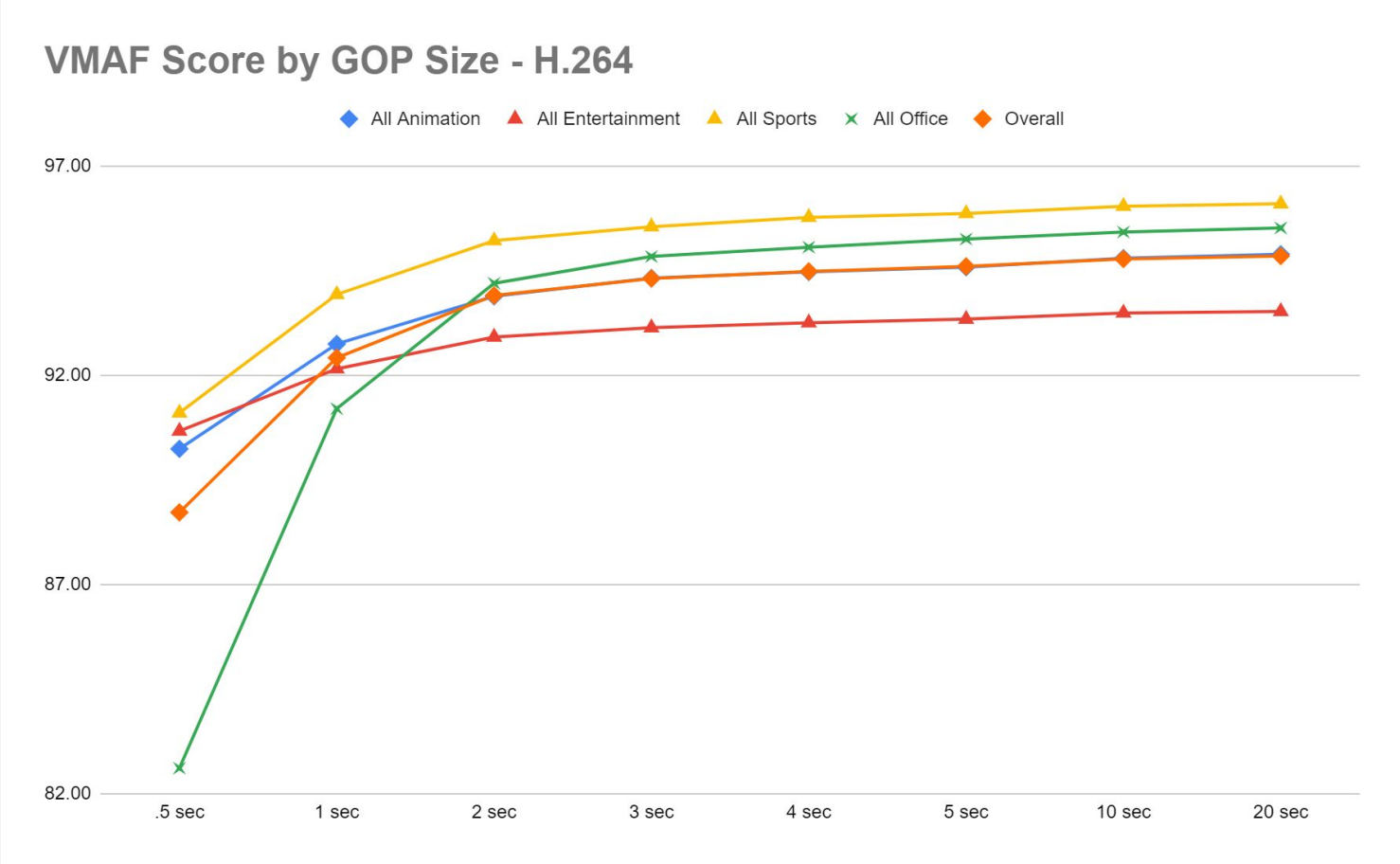
H.264 - All Content Types

- Animation
- Entertainment
- Sports
- Office

Office clips showed the most degradation with low GOP sizes, followed by animation.

Entertainment content was surprisingly resilient at low GOP sizes.

Overall the difference between 10 and 20 seconds was minimal.



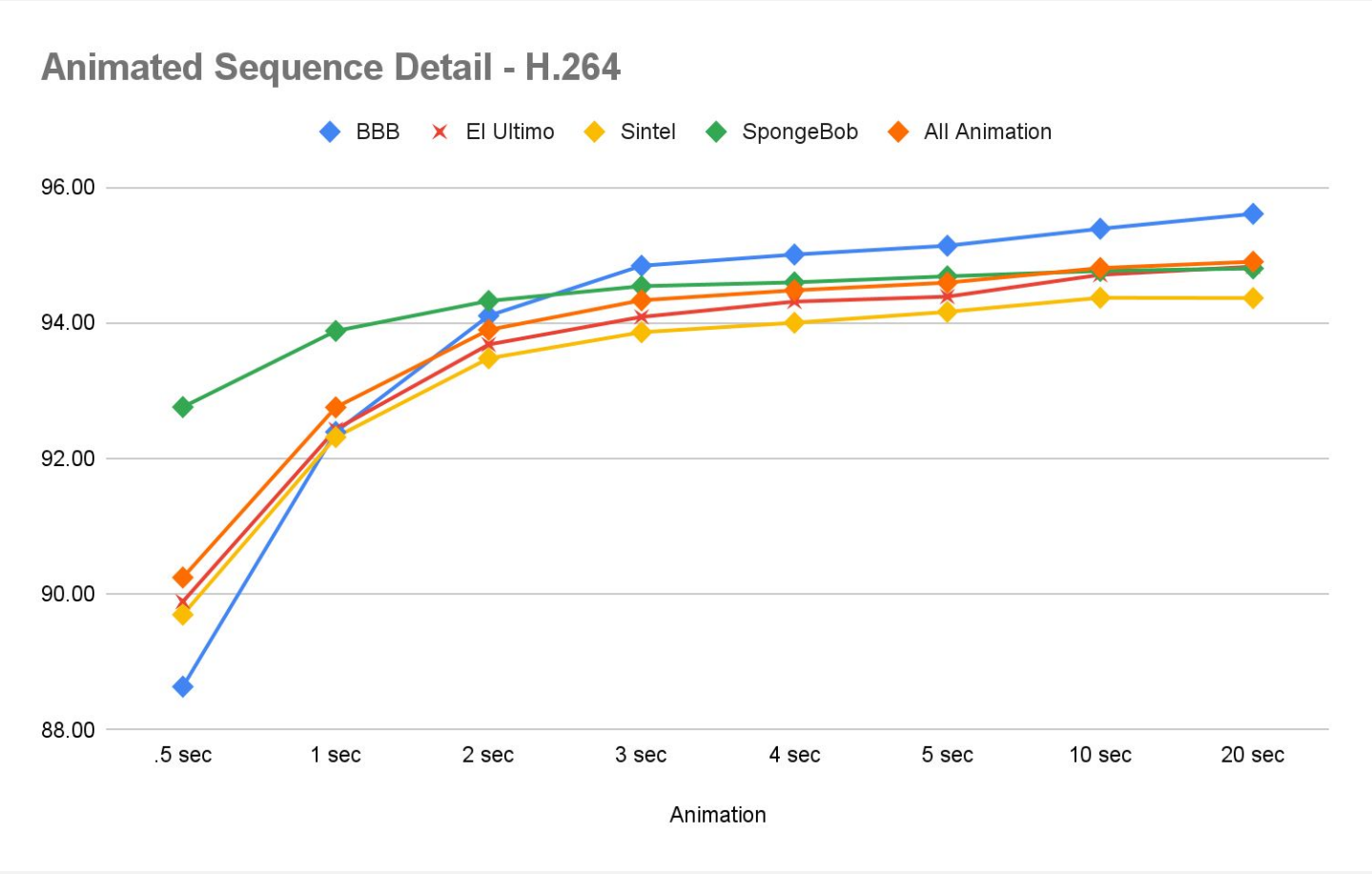
Overall - H.264	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
All Animation	90.25	92.75	93.90	94.33	94.48	94.59	94.81	94.90
All Entertainment	90.67	92.16	92.92	93.14	93.26	93.35	93.50	93.53
All Sports	91.11	93.94	95.23	95.56	95.78	95.88	96.05	96.11
All Office	82.61	91.21	94.21	94.85	95.07	95.26	95.43	95.53
Overall	**88.73**	**92.42**	**93.92**	**94.32**	**94.49**	**94.61**	**94.79**	**94.86**
Delta from Max	**6.13**	**2.43**	**0.94**	**0.54**	**0.37**	**0.25**	**0.07**	**0.00**

H.264 - Animation

Big Buck Bunny kept improving from 10 to 20 seconds but most other files were relatively static.

Spongebob was the most real-worldish in content, and like the entertainment clips, wasn't that impacted by very short GOP sizes.

Others clumped together. I would avoid very low GOP sizes with animated content.

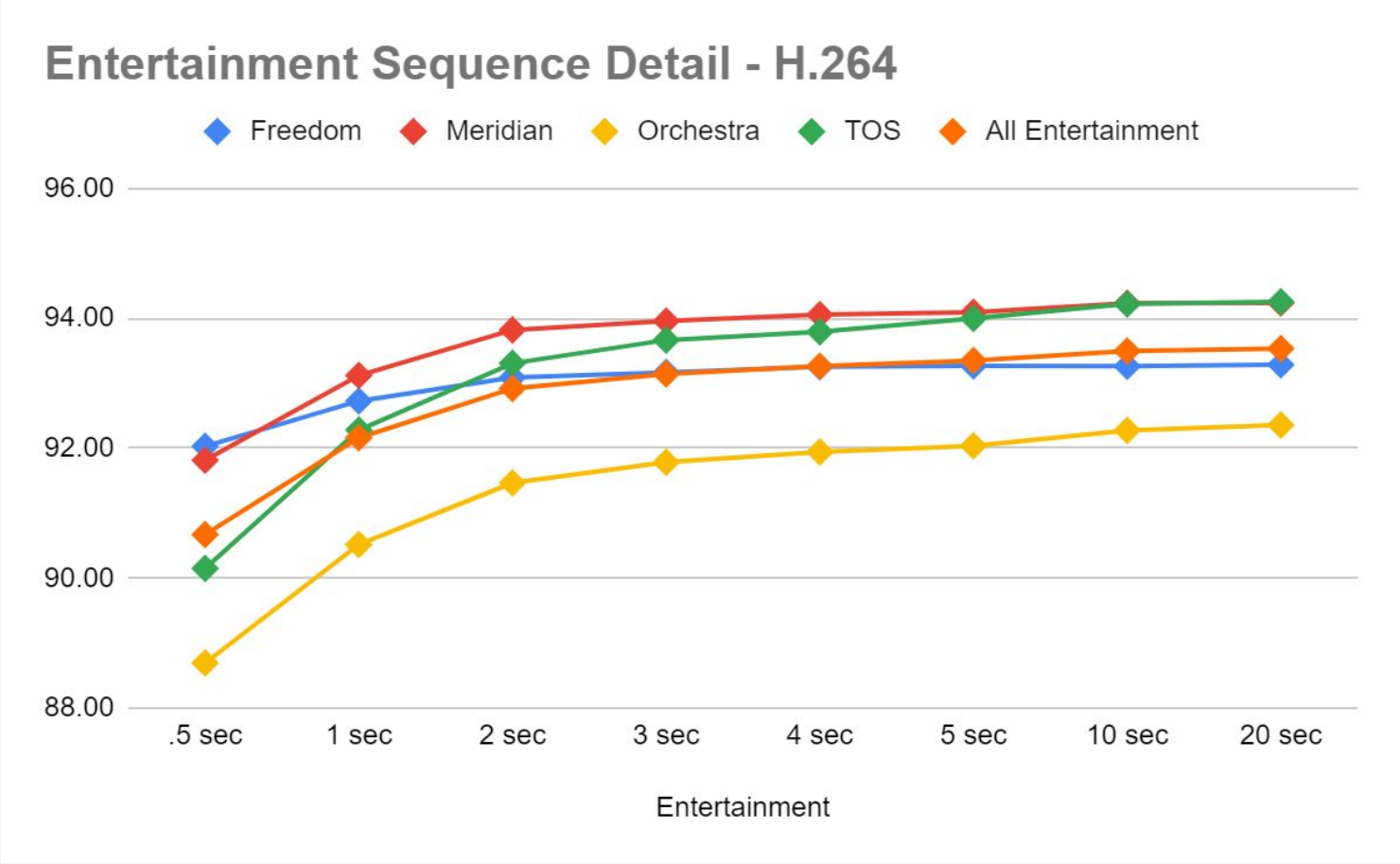


Animation - H.264	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
BBB	88.64	92.39	94.10	94.84	95.01	95.13	95.39	95.61
El Ultimo	89.89	92.43	93.68	94.09	94.31	94.39	94.71	94.83
Sintel	89.70	92.31	93.48	93.86	94.00	94.16	94.37	94.36
SpongeBob	92.76	93.88	94.32	94.54	94.60	94.68	94.76	94.80
All Animation	**90.25**	**92.75**	**93.90**	**94.33**	**94.48**	**94.59**	**94.81**	**94.90**
Delta from Max	**4.65**	**2.15**	**1.00**	**0.57**	**0.42**	**0.31**	**0.09**	**0.00**

H.264 - Entertainment

These clips showed the least impact from very small GOP sizes and the lowest overall delta between the highest and lowest scores with a fairly consistent pattern except for Freedom, which is a concert video clip.

There was a substantial difference between one and two seconds, much less between two and three seconds.



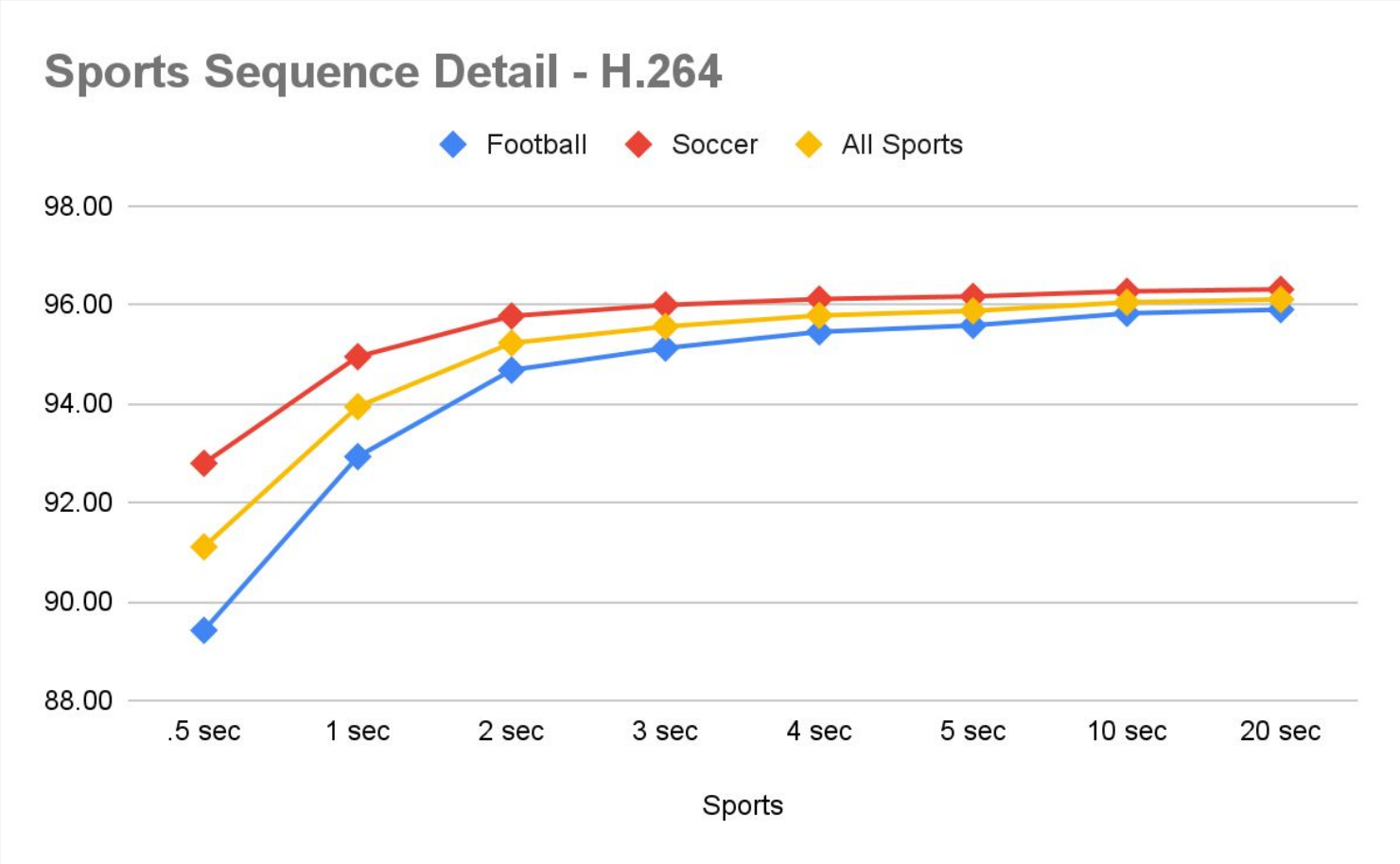
Entertainment - H.264	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
Freedom	92.03	92.73	93.09	93.17	93.26	93.27	93.26	93.28
Meridian	91.81	93.12	93.82	93.96	94.06	94.09	94.23	94.23
Orchestra	88.69	90.52	91.47	91.78	91.94	92.04	92.27	92.36
TOS	90.15	92.28	93.31	93.66	93.79	94.00	94.22	94.25
All Entertainment	**90.67**	**92.16**	**92.92**	**93.14**	**93.26**	**93.35**	**93.50**	**93.53**
Delta from Max	**2.86**	**1.37**	**0.61**	**0.39**	**0.27**	**0.18**	**0.04**	**0.00**

H.264 - Sports

Sorry just two clips here. Football is a tough clip, soccer slightly easier.

I recognize that most sports broadcasts are live, which means single pass, but I couldn't resist testing these iconic (for me) clips.

I used two-pass encoding for these tests, but beware of small GOP sizes used for live transcoding in an effort to reduce latency; it will reduce quality. Even the difference between 1 and 2 seconds was meaningful for the football clip.

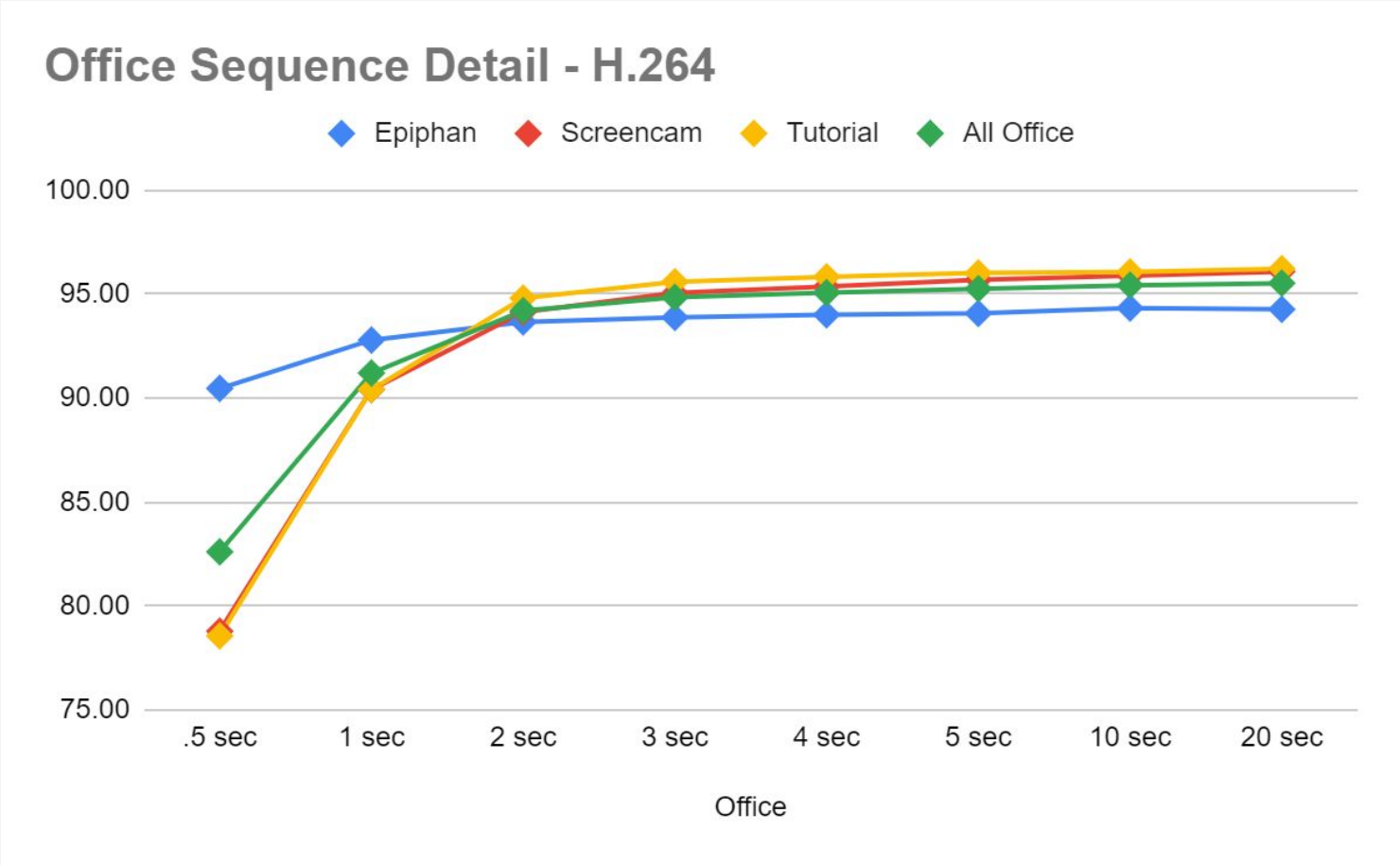


Sports - H.264	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
Football	89.42	92.93	94.68	95.12	95.45	95.58	95.83	95.90
Soccer	92.79	94.95	95.77	95.99	96.11	96.17	96.27	96.31
All Sports	**91.11**	**93.94**	**95.23**	**95.56**	**95.78**	**95.88**	**96.05**	**96.11**
Delta from Max	**5.00**	**2.17**	**0.88**	**0.55**	**0.32**	**0.23**	**0.06**	**0.00**

H.264 - Office

Epiphan is mixed screencam with live video; it performed like the entertainment videos.

Screencam is Camtasia based, and Tutorial is PowerPoint slides with video. These two synthetic clips showed extreme degradation at smallish GOP sizes; if distributing on a corporate network, go with GOP sizes of at least three seconds, preferably four or five.



Office - H.264	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
Epiphan	90.46	92.79	93.66	93.88	94.01	94.07	94.34	94.27
Screencam	78.79	90.41	94.14	95.06	95.36	95.69	95.89	96.09
Tutorial	78.57	90.41	94.81	95.60	95.84	96.02	96.07	96.23
All Office	**82.61**	**91.21**	**94.21**	**94.85**	**95.07**	**95.26**	**95.43**	**95.53**
Delta from Max	**12.92**	**4.32**	**1.32**	**0.68**	**0.46**	**0.27**	**0.10**	**0.00**

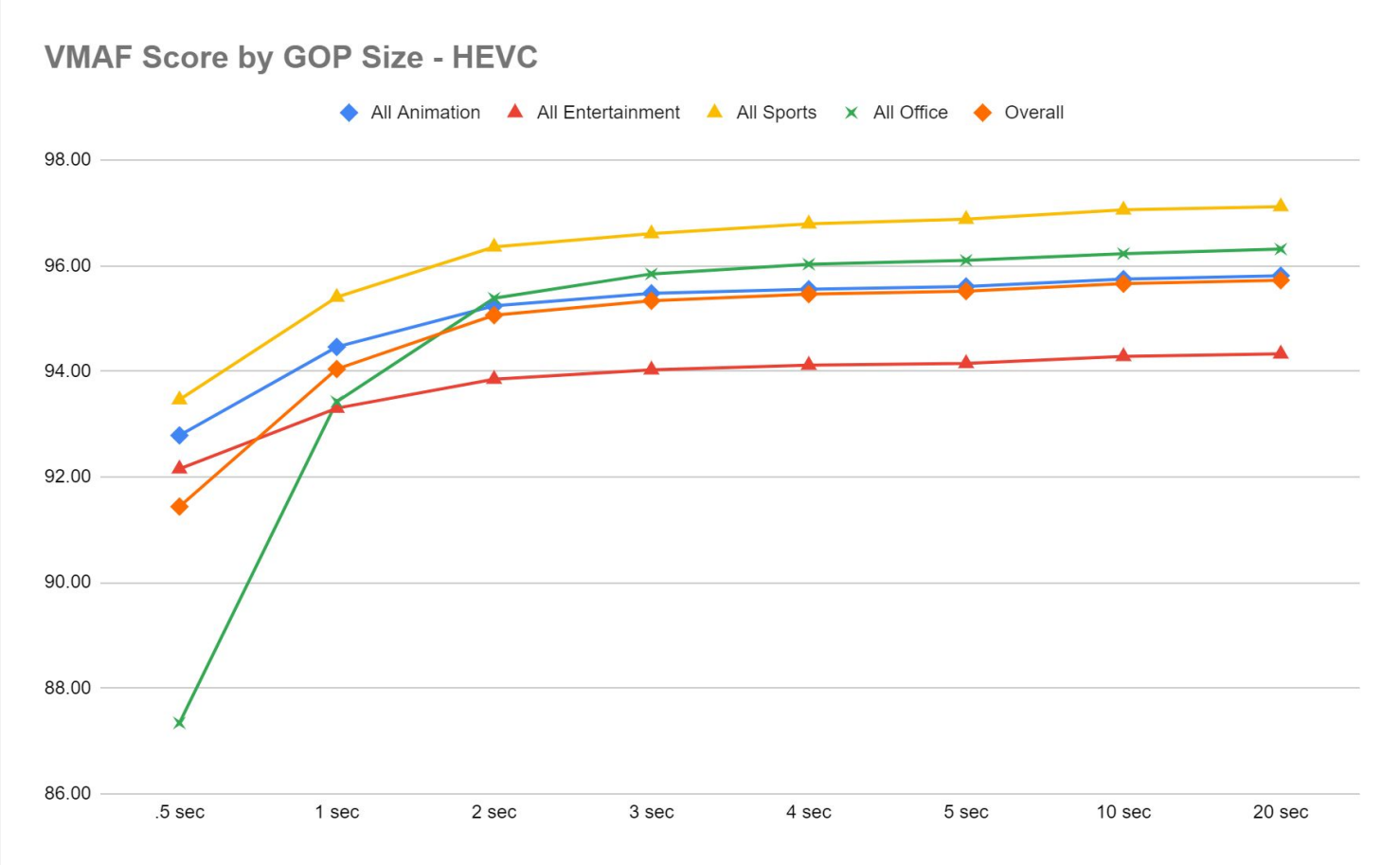
HEVC - All Content Types

- Animation
- Entertainment
- Sports
- Office

As with H.264, office clips showed the most degradation with low GOP sizes, with entertainment the most resistant.

Unlike H.264, which showed significant degradation with low GOP sizes, animation was more resistant with HEVC.

Overall, HEVC proved more resilient to small GOP quality issues, with a max delta of 4.29 compared to 6.13 for H.264. You'll see a comparison a few slides down.

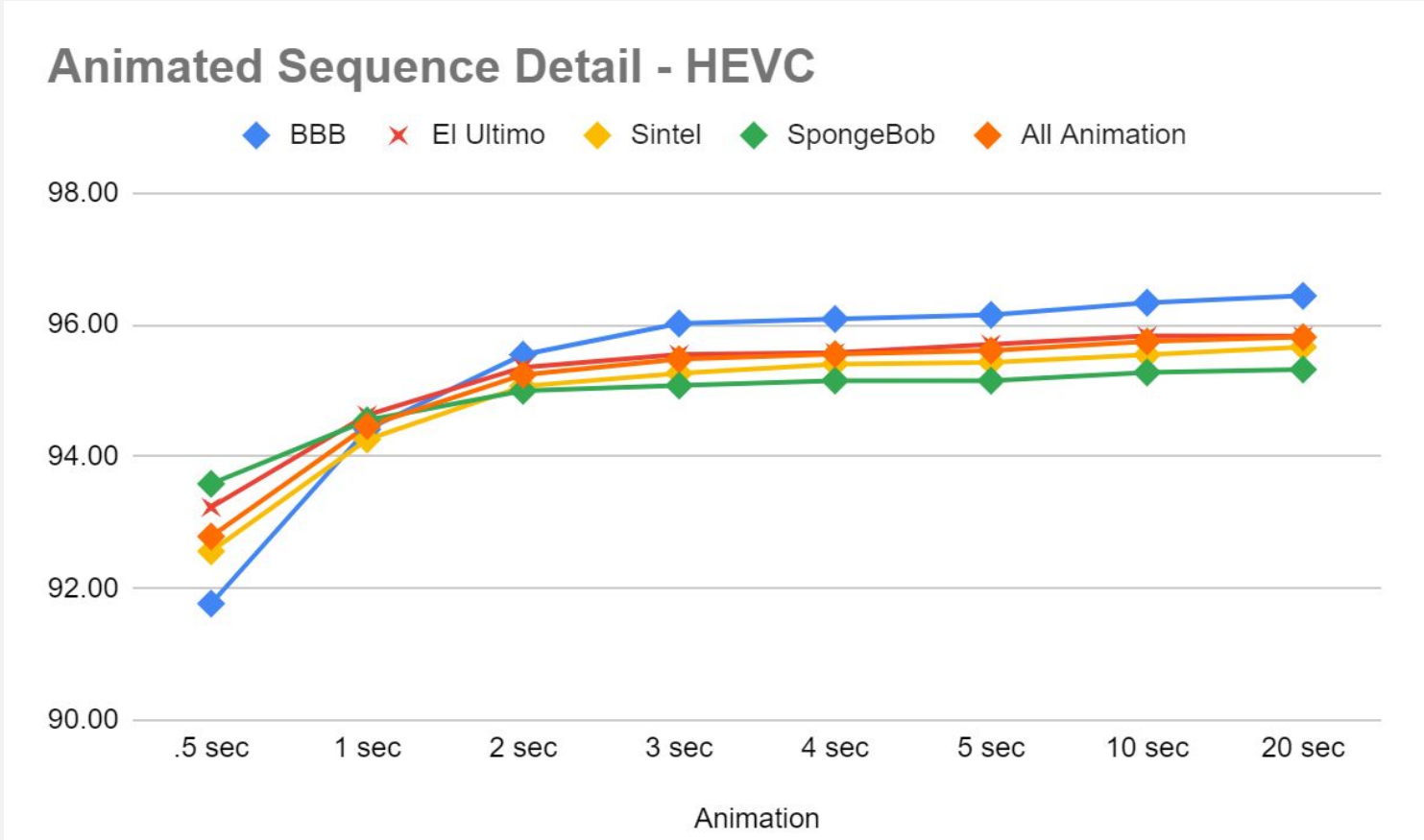


Overall - HEVC	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
All Animation	92.79	94.46	95.24	95.48	95.56	95.61	95.75	95.81
All Entertainment	92.15	93.30	93.85	94.03	94.12	94.15	94.28	94.33
All Sports	93.46	95.41	96.36	96.61	96.80	96.88	97.06	97.12
All Office	87.34	93.43	95.39	95.85	96.03	96.10	96.23	96.32
Overall	**91.44**	**94.04**	**95.06**	**95.34**	**95.46**	**95.52**	**95.66**	**95.73**
Delta from Max	**4.29**	**1.68**	**0.66**	**0.39**	**0.26**	**0.21**	**0.06**	**0.00**

HEVC - Animation

As with H.264, Big Buck Bunny showed the highest differential and Spongebob the least.

Almost all the clips showed little improvement between three to five seconds, with slight improvement through to 20 seconds.

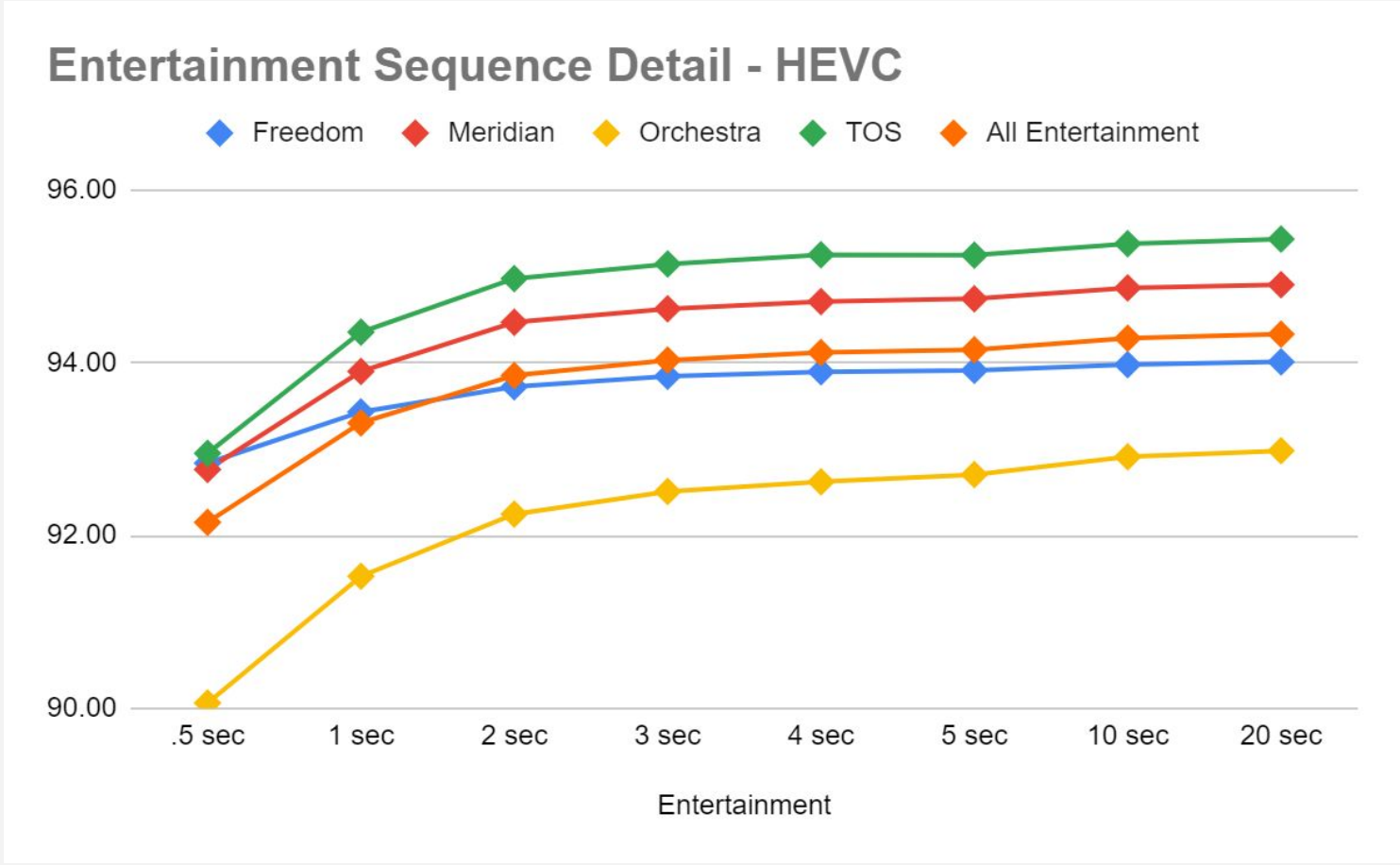


Animation - HEVC	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
BBB	91.77	94.41	95.55	96.02	96.09	96.15	96.34	96.44
El Ultimo	93.23	94.63	95.35	95.55	95.58	95.70	95.83	95.83
Sintel	92.56	94.26	95.07	95.26	95.40	95.43	95.54	95.66
SpongeBob	93.59	94.55	95.00	95.08	95.15	95.15	95.28	95.32
All Animation	**92.79**	**94.46**	**95.24**	**95.48**	**95.56**	**95.61**	**95.75**	**95.81**
Delta from Max	**3.03**	**1.35**	**0.57**	**0.34**	**0.26**	**0.20**	**0.07**	**0.00**

HEVC - Entertainment

Again, entertainment clips showed the least impact from very small GOP sizes and the lowest overall delta between the highest and lowest scores with a fairly consistent pattern except for Freedom, which is a concert video clip. I'm not sure why.

Orchestra which is a tough clip with lots of instruments and sharp edges, showed the greatest delta, but it looks worse than it is because the VMAF scores are about a point lower than the others. My bad on that one.

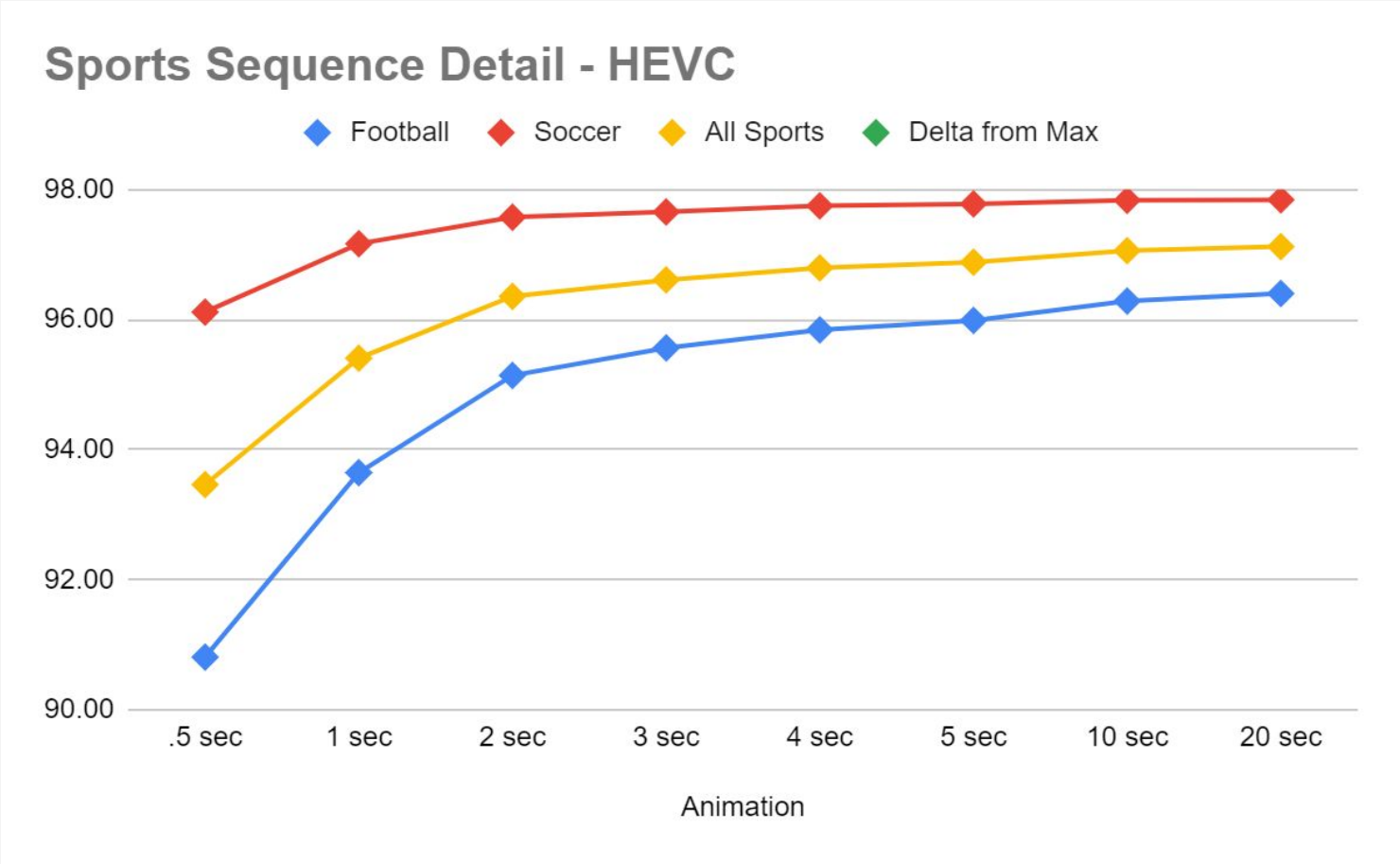


Entertainment - HEVC	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
Freedom	92.84	93.43	93.72	93.84	93.89	93.91	93.98	94.01
Meridian	92.76	93.90	94.47	94.62	94.71	94.74	94.87	94.90
Orchestra	90.06	91.53	92.25	92.51	92.62	92.71	92.91	92.98
TOS	92.95	94.35	94.97	95.14	95.25	95.24	95.38	95.43
All Entertainment	**92.15**	**93.30**	**93.85**	**94.03**	**94.12**	**94.15**	**94.28**	**94.33**
Delta from Max	**2.18**	**1.03**	**0.48**	**0.30**	**0.21**	**0.18**	**0.05**	**0.00**

HEVC - Sports

Football showed very significant overall delta (~5.6). I'm not sure why soccer showed much less delta (~1.7) since it's an equally tough clip.

Again, be careful with GOP size when encoding for low latency. Small GOP sizes have an outsized impact with sports clips.



Sports - HEVC	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
Football	90.81	93.65	95.14	95.56	95.84	95.99	96.28	96.40
Soccer	96.12	97.17	97.57	97.66	97.75	97.78	97.83	97.84
All Sports	**93.46**	**95.41**	**96.36**	**96.61**	**96.80**	**96.88**	**97.06**	**97.12**
Delta from Max	**3.66**	**1.72**	**0.76**	**0.51**	**0.33**	**0.24**	**0.06**	**0.00**

HEVC - Office

Screencam and Tutorial are predominantly synthetic content, and really suffered from low GOP sizes. Epiphan was mixed, and fared much better.

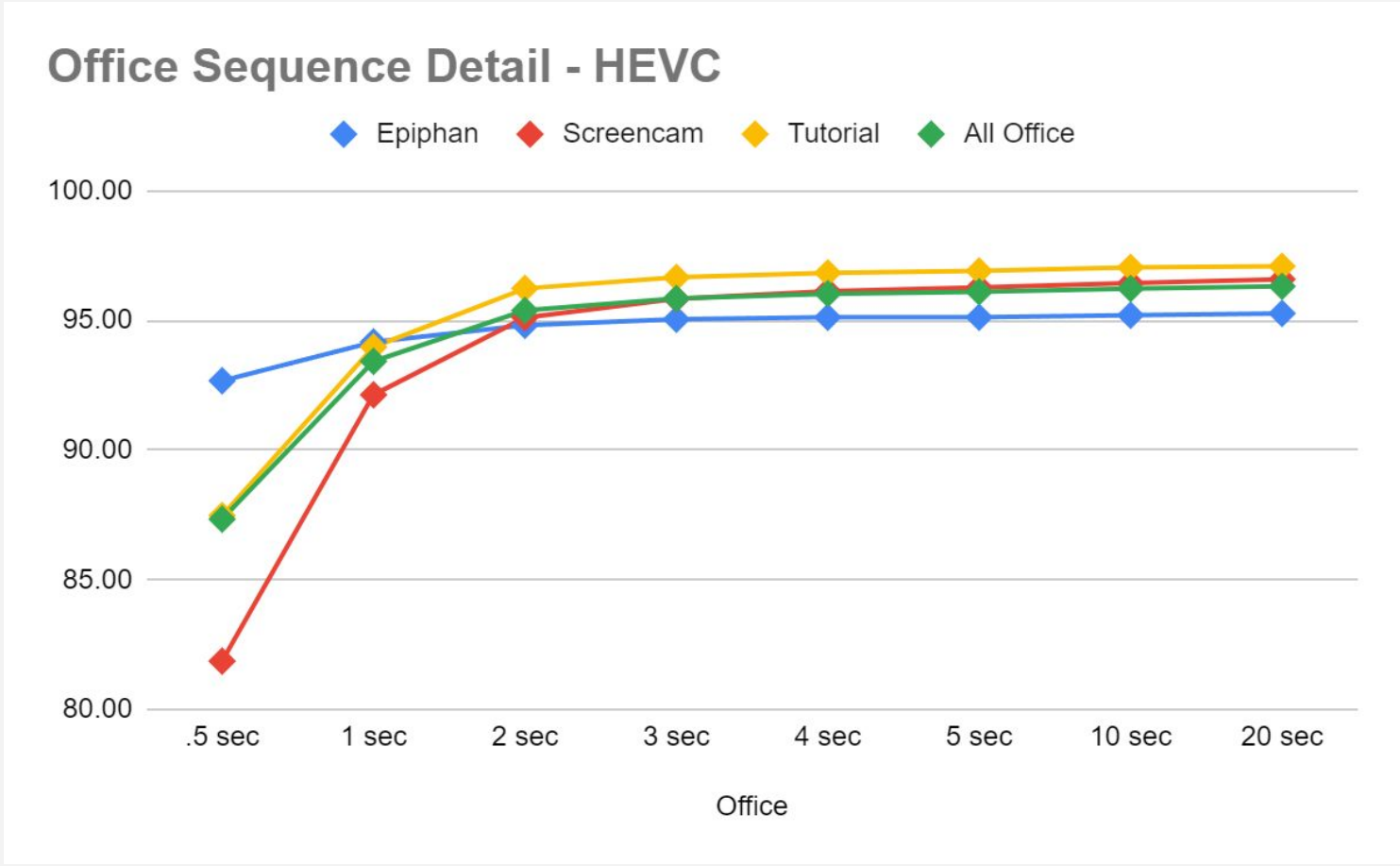
If you're encoding screencams with x265, it may pay to experiment. Here's some screencam encoding advice from Ben Waggonner (who heads up encoding for Amazon Prime) from here.

Enabling --tskip may improve the compression efficiency of screen content (aka: text on a screen)

And it is losslessly acquired (so zero noise), --cu-lossless might even help sometimes.

I'd also couple a high --bframes with the high --keyint. Random access time for this kind of content is proportional to keyint/bframes, and bframes are particularly efficient with this kind of content.

I'd also want --tu-inter and -tu-intra to be 4. Lots of fine sharp details in screen recordings, so you want to recurse tu size down as low as possible.

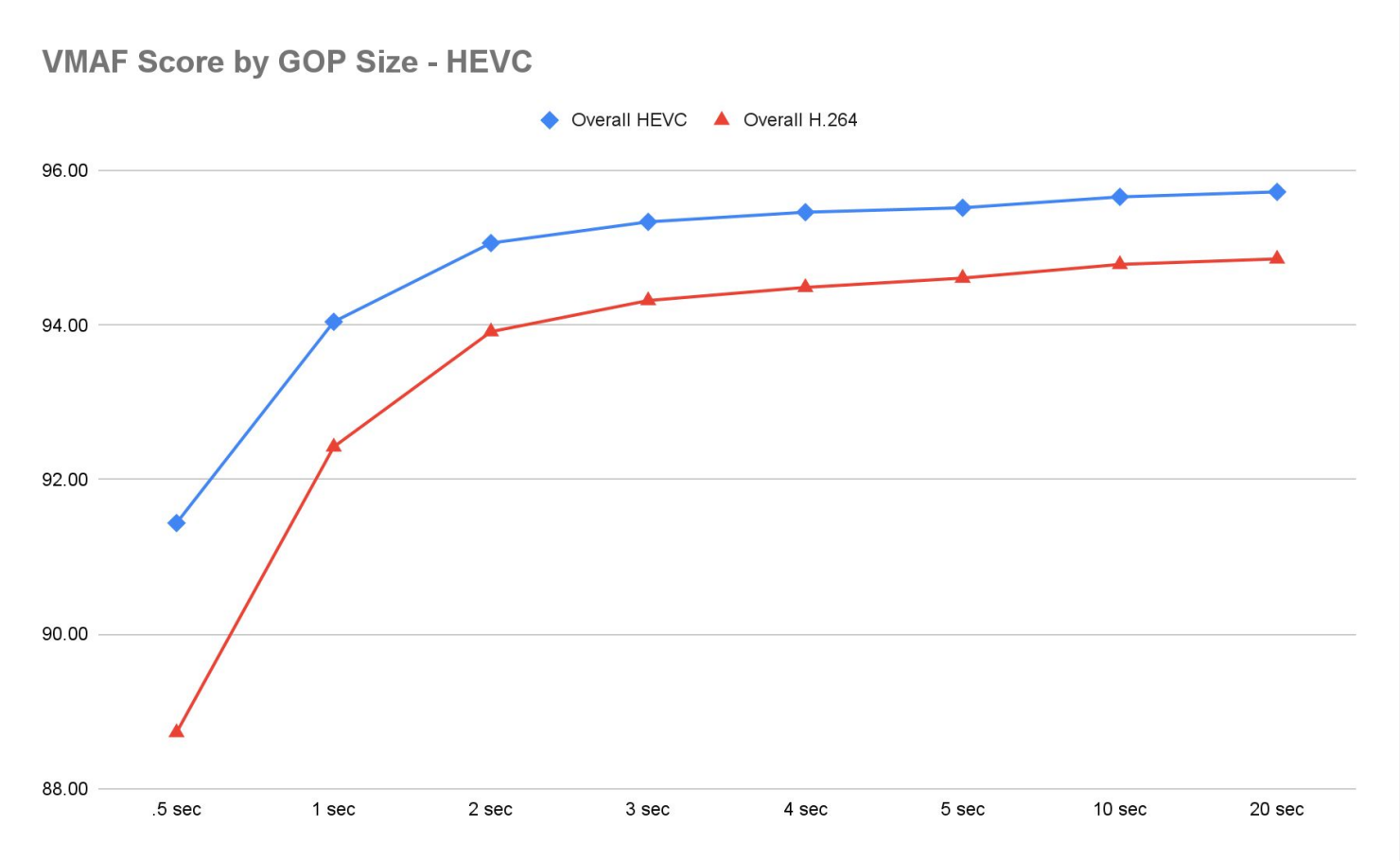


Office - HEVC	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec
Epiphan	92.67	94.17	94.81	95.05	95.12	95.12	95.20
Screencam	81.87	92.13	95.11	95.83	96.13	96.28	96.44
Tutorial	87.48	93.98	96.23	96.66	96.84	96.91	97.05
All Office	**87.34**	**93.43**	**95.39**	**95.85**	**96.03**	**96.10**	**96.23**
Delta from Max	**8.98**	**2.89**	**0.93**	**0.47**	**0.29**	**0.21**	**0.09**

H.264 vs HEVC

Though I targeted the same VMAF level for both sets of encodes, HEVC obviously crept a bit higher. Still, from the max delta score, it's clear that HEVC better resists degradation from small GOP sizes.

With both, going beyond five second GOP sizes delivers irrelevant benefits.



GOP Size	.5 sec	1 sec	2 sec	3 sec	4 sec	5 sec	10 sec	20 sec
Overall HEVC	91.44	94.04	95.06	95.34	95.46	95.52	95.66	95.73
Max delta - HEVC	4.29	1.68	0.66	0.39	0.26	0.21	0.06	0.00
Overall H.264	88.73	92.42	93.92	94.32	94.49	94.61	94.79	94.86
Max delta - H264	6.13	2.43	0.94	0.54	0.37	0.25	0.07	0.00

Summary and Conclusion

These results will surprise very few experienced compressionists, but you shouldn't make recommendations in courses, books, magazines, and blog posts without having the data, and now I do.

Interestingly, many compression grey hairs who lived through the cable TV days used a GOP size of 15 frames for their streaming files because that was common for MPEG-2. Data contained herein should discourage that.

From a content type perspective, real world videos seemed the least impacted by very short GOP sizes, while synthetic videos like animations and particularly screencams, were most affected.

Otherwise, those tempted to drop the GOP size in low latency applications should avoid going under 2 seconds or so. In most cases, this won't impact latency, since schema operation isn't based on I-frames.

If you have any questions or comments, please drop me a note at janozer@streaminglearningcenter.com.

STREAMING LEARNING CENTER