Can I add a VR flow? On the maximum capacity of 5G to support 360° video

Pablo Serrano*, Antonio Virdis†, Francesco Gringoli‡, Marco Gramaglia*
*University Carlos III of Madrid, Spain
†Universita di Pisa, Italy
‡University of Brescia/CNIT, Italy

***Abstract*—360° video is gaining popularity thanks to the increasing prevalence of Virtual Reality (VR) devices. This has motivated novel approaches to improve the efficiency of 360° video transmission, with techniques that range from content distribution techniques (using edge servers) to the so-called viewport or tile prediction. In this paper, we take a different approach and study the maximum capacity of a 5G cell to support 360° video flows. Specifically, we provide a methodology to estimate the upper bound on the number of simultaneous 360° video flows that can fit the cell. To this aim, we first define different quality profiles for video transmission, which are based on subjective video quality metrics. This upper bound is calculated as a function of the quality profile of the video transmission. We also analyze the impact of different parameters on the results, including scenarios with interfering cells. Our results quantify the ability of 5G cells to support VR traffic and the impact of the type of video or the quality of experience on performance.**

I. Introduction

The advent of 5G technology marks a significant milestone in the evolution of wireless communication systems. Characterized by its high-speed data transfer, reduced latency, and increased connectivity, 5G is poised to revolutionize various sectors including healthcare, automotive, and smart cities. Its deployment involves upgrading existing cellular infrastructure and installing new cell towers equipped with advanced hardware to support the 5G spectrum.

Virtual reality (VR) technology, especially in the context of the metaverse, poses significant challenges for 5G networks. The metaverse depends extensively on VR technology to deliver immersive and interactive experiences that closely simulate real-world environments. Achieving the high bandwidth and low latency essential for an uninterrupted and high-quality experience is critical, yet it challenges the existing capabilities of 5G cells. The capacity of current 5G networks may restrict the number of users who can simultaneously engage in these environments.

When assessing the impact of network capacity on VR experiences, Quality of Experience (QoE) metrics such as the Video Multimethod Assessment Fusion (VMAF), play a crucial role. VMAF [1], developed by Netflix, is a perceptual video quality measurement that integrates various quality metrics to predict subjective video quality more accurately. Although not initially designed for 360° video content, recent studies have demonstrated that VMAF performs adequately well with this type of content [2]. In the context of VR, VMAF helps in quantifying the user's perceived quality by evaluating factors such as resolution, buffering events, and bitrate fluctuations, which are directly influenced by the network's capacity. The effective use of VMAF in VR scenarios allows researchers and network providers to understand how bandwidth limitations and network congestion can degrade the visual fidelity and overall immersive experience of VR content. This understanding is vital for optimizing network resources, particularly in 5G environments, to ensure that the high data requirements of VR are met without compromising the user's immersive experience. By leveraging VMAF and similar QoE metrics, stakeholders can make informed decisions about network planning and management, aiming to enhance the overall quality and accessibility of VR services.

In this paper, we design a QoE-driven approach to quantifying the capacity of 5G to support VR services. Our approach is complementary to recent works such as [3], which in contrast to QoE, focuses on high-level Key Quality Indicators (KQI) such as throughput, delay, frame rate, or stall events. We aim to provide providers with a methodology and a tool to understand the trade-off between the number of served users and the experienced quality, in this way supporting e.g. the ability to assess if a new VR flow can be added to the cell without affecting the QoE of existing flows. To this aim, we first define a video transmission model, based on different "focus areas" as motivated by perceptual models. We then collect several 360° video traces, composed of videos of different natures, and emulate different transmission configurations following the above model, taking into account the impact of the transmission rate on the QoE. Based on these results, we finally present a capacity study about the number of users that can be served depending on the video considered, transmission profiles, and other conditions of the scenario.

The rest of the paper is structured as follows. In Section II we describe the adopted video transmission model, which consists of two flows of different quality; in Section III we detail our QoE-driven methodology to determine the transmission profiles considered; in Section IV we present our capacity study, describing the 5G capacity model and analyzing for different scenarios how many 360° video flows can be supported; finally, we summarize the paper and provide some ideas for future work in Section V.

II. Video transmission model

Following the classical definition, the *visual field* of the human eye is the area visible during stable fixation of the eyes (i.e., eye movements are excluded in this definition), specified in degrees of visual angle [4]. It can be decomposed into a horizontal visual field and a vertical visual field. The perception of the eye changes depending on the angle range around the eye's focal point [5]: The central vision, approx. a 60° angle (i.e., ±30°) from the focal point) is the most sensitive to the details, while the peripheral vision (an extra ±30° angle from the central vision) is less sensible to static object recognition but still receptive to movements.

The *Field of View (FoV)* is the corresponding definition for optical devices (such as VR headsets) when eye movements are allowed. Formally speaking, the FoV is the area of all points on a unit sphere around the human eye that correspond to directions that end up on a screen. In other words, the visual field tells us what the eye can potentially see (and how well it perceives it) whereas the FoV describes the amount of "angle" of vision of a human eye that is shown on the VR screen.

A prevalent technique for VR transmission involves segmenting each image into numerous rectangular segments, commonly referred to as tiles. This method hinges on a selective transmission strategy, where only a crucial subset of these tiles—those encompassing the viewer's current FoV—is actively transmitted. Recent works that leverage deep learning solutions, trained on the device onboard sensors such as gyroscopes and accelerators show promising results for such kinds of strategies [6], which also require overwriting of the playout buffer when sudden changes are detected [7].

In this paper, to investigate the capacity limits of 5G to transmit 360° video we postulate a simplified transmission model, wherein the video is bifurcated into two distinct streams. This model is inspired by YBVR's (https://ybvr.com) proprietary technology which applies an optical deformation that maintains maximum resolution at the FoV and reduces it outside of that field. The first stream, which we call the "front" stream, corresponds to the viewer's FoV and is transmitted at the *best* quality to ensure a great viewing experience. The second stream, which we call the "full" stream, encompasses the remainder of the video, which is transmitted at an *adequate* quality that balances bandwidth utilization and overall video experience (which ultimately impacts the requirements on the tile prediction algorithm).

To devise the video resolution of the above streams, we considered the well-known HTC Vive visor (https://www.vive.com/us/) and different 8K 360° videos from YouTube as video sources. The former provides a horizontal and vertical FoV of 108° and 111°, respectively (note that other visors in the marker, e.g., the Oculus Meta Quest Pro, share similar parameters), while the latter uses a custom resolution of 3840p, namely 7680x3840, which corresponds to 360° in the horizontal axis and 180° in the vertical axis. Following these values, we compute the required resolution of the front stream by proportionally cropping a portion of the video corresponding to the FoV of the HTC Vive. In the following section, we describe the videos used in our analysis and elaborate on what constitutes 'best' and 'adequate' quality in this context.

III. QoE-driven definition of profiles

In the above, we have defined a transmission model based on two flows, one for the front video (also known as field of view), and another one for the rest of the video, i.e., the full view (a.k.a. equirectangular view). To define the different videos and transmission profiles and requirements, we next: (1) describe the videos used in the paper, (2) discuss the criteria to determine a good and adequate quality, (3) design a methodology to map the quality to the required transmission rate, and (4) introduce the different transmission profiles considered in the rest of the paper.

A. Videos considered

Although our motivation is live streaming, for repeatability we decided to study the videos in the YouTube *AirPano VR* channel [8]. To simplify the presentation of the results while ensuring the conclusions are broadly applicable, we selected three videos as a representative sample of the range of behavior seen. Since these videos encapsulate a broad range of performance behaviors observed across the heterogeneous dataset considered, it allows for the generalization of findings, suggesting that the insights derived from these selected videos are likely applicable to other videos beyond the ones specifically analyzed. We downloaded local copies for later analysis on a workstation powered by an 18-core Intel i9-10980XE CPU: to this end, we used `yt-dlp`, an open-source python script that allows us to specify the desired resolution, which in this case is set to 8K. All videos are originally encoded with the AV1 codec and, because of the equirectangular format, their resolution is 7680x3840 pixels (as confirmed by `ffmpeg`). We properly cut each video to eliminate the opening credits and select a 10 s scene from each video. We describe the three scenes next:

Everest: The video depicts a helicopter's takeoff from a base camp. The first 5 s are recorded by a ground camera, while the last 5 s are captured by a camera on a pole attached to the helicopter.

Caribbean: The video captures boats sailing on the Caribbean Sea. It features short waves and small plant-covered islands in a sunlit day.

Buffaloes: A static camera films a nearly-still herd of buffaloes in a snowy winter prairie. In the final 3 s, one buffalo attacks another one.

To provide an idea about the content and dynamics of each video, we provide in Fig 1 several stills from each of the videos. Each row corresponds to a different video, which are (from top to bottom): Everest, Caribbean, and Buffaloes. For each video, we provide in the leftmost picture the full (equirectangular) view at t = 5 s, highlighting in red the area that corresponds to the front view. The stills corresponding to these front views at t = 0 s, t = 5 s, and t = 10 s are provided in the three rightmost pictures in that order.

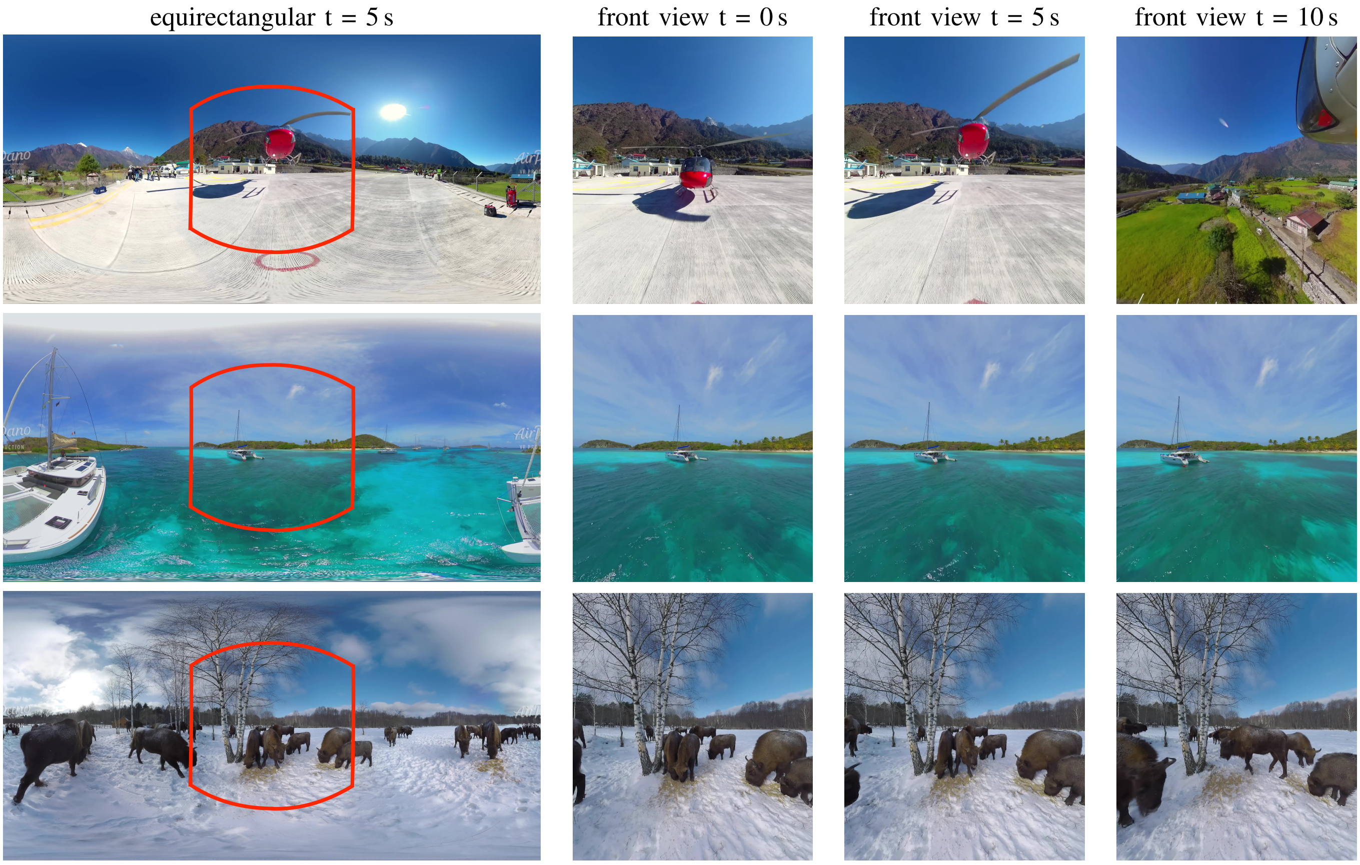


Fig. 1: Stills extracted from the three videos. From top to bottom: Everest, Caribbean, Buffaloes. Left: full view at t = 5 s, the red freeform corresponds to the front view. Right: front views at t = 0 s, 5 s, and 10 s.

B. QoE metric: VMAF

We rely on Netflix's VMAF metric (as we previously did in [9]) as, unlike SSIM which focuses primarily on changes in structural information, or PSNR, which measures pixel-level differences, it integrates multiple quality metrics and machine learning algorithms to evaluate video quality. This holistic approach allows VMAF to consider a range of factors that affect human perception, including texture, motion, and spatial complexities, providing a more comprehensive and accurate reflection of perceived video quality. Furthermore, VMAF is trained using a large dataset of human-rated video samples, enabling it to predict subjective video quality with greater precision. This makes VMAF particularly valuable in scenarios where user experience is paramount.

VMAF scores range from 0 to 100, with scores closer to 100 indicating excellent quality, but the precise interpretation of scores can vary depending on factors like the viewing conditions. According to some sources [10], a score of 20 is "bad", a score of 100 is "excellent", while a score of 70 can be interpreted as somewhere between "good" and "fair" by an average viewer; other sources [11] claim that any value exceeding 95 is "wasting bandwidth" and that for premium content, the target should be between 93 and 95, while for user-generated content, scores between 84 and 92 are "acceptable". Based on these values, we decided to set the following thresholds:

- Front view, requiring the best quality: this translates into a VMAF score of 95.
- Full view (rest of the video), requiring an adequate quality: a VMAF score between 70 and 92.

C. Methodology

As mentioned above, our goal is to determine the VMAF corresponding to different video encoding rates for (1) full view (i.e., the entire equirectangular 360 frame), and (2) the front view that appears on the headset, i.e., FoV. We first describe how we achieved the first goal, and then explain how to adapt the same methodology for the second goal. We note that, during our tests, we chose Advanced Video Coding (AVC1) over AOMedia Video (AV1) to generate the videos because of its lower computing power requirements, making it suitable for real-time operations.[1]

We first trim the video to extract the desired 10 s scene. To this end, we skip an integer number of Group-of-Pictures (GOP) from the beginning until the starting time of the scene and then save an integer number of GOPs that correspond to at least 10 s (note that different videos use different GOP lengths). Working with GOPs allows us not to transcode the videos to extract the scenes, hence preserving the same

[1] Should AV1 encoding become fast enough, similar conclusions might be drawn in relative terms, i.e., the ratio between the data rates of the equirectangular and the front view videos should be comparable.

quality as in the original content. We then generate two reference points: (a) we explode with ffmpeg each video in uncompressed yuv format inside an AVI container, which will serve as the starting point for each encoding using a different rate; and (b) we save the uncompressed content as a sequence of yuv pictures in a binary file using a gstreamer pipeline, which will serve as the reference for VMAF computation.

Based on the above, we repeat for each considered encoding rate the following steps:[2]

1) We re-encode the uncompressed yuv AVI file with ffmpeg using the AVC1 codec from the libx264 library. We use two passes to obtain a given rate.
2) We explode the re-encoded video into a new sequence of yuv pictures using the same gstreamer pipeline as before;
3) We compute the VMAF score using this last sequence and the reference yuv sequence generated in the step (b) above.

To compute the corresponding VMAF scores for the different encoding rates for the front view, we follow the same methodology but starting from a video containing just the view that would appear in the headset. We generate a new video with ffmpeg starting from the uncompressed AVI file but applying the v360 filter of ffmpeg to extract the central view. We then repeat the steps above to compute the VMAF scores corresponding to the encoding rates. We next discuss the results obtained and based on these the design of the different transmission profiles.

D. Considered transmission profiles

We depict in Fig 2 the resulting VMAF scores for the different videos and encoding rates considered: Fig. 2a illustrates the results corresponding to the full view (equirectangular videos), i.e., the whole 360 frame, while Fig. 2b illustrates the results corresponding to the front view (note the different scale in the x-axis).

According to the results, the “Caribbean” video is the most demanding to encode in its full view (Fig. 2a), since it requires an encoding rate above 25 Mb/s to achieve a VMAF score above 70; in contrast, it is the least demanding in the its front view (Fig. 2b), reaching VMAF scores above 80 or 90 for transmission rates smaller than for the other two videos. This can be explained because of the prevalent sea waves in the 360 format, which are much less noticeable in the front view (see Fig. 1). The “Everest” video shows the qualitative opposite behaviour: the full video vision reaches a VMAF score of 90 at approx. 25 Mb/s, but the front view requires the largest transmission rate to reach a VMAF score of 80. This can be explained because of the change of camera during the video that heavily affects the front view (see Fig. 1, first row, at t=5 s and t=10 s). Finally, the “Buffaloes” video roughly falls between these two behaviors.

To determine the different transmission profiles, note that in Section III-B we discussed a set of target values for VMAF scores. Following these, depending on the considered vision, we define the following.

- Front view: to provide the best possible quality, a VMAF score of 95 is required. According to the results in Fig. 2b, this requires a transmission rate of 4.2 Mb/s, 8 Mb/s, and 6.5 Mb/s for the Caribbean, Buffaloes, and Everest videos, respectively.
- Full view / equirectangular video: to provide adequate quality, the VMAF scores should follow between 70 and 92. This translates in the following transmission rate ranges: {26–70} Mb/s for the Caribbean video, {11–70} Mb/s for the Buffaloes video, and {10–31} Mb/s for the Everest video.

Following these transmission rate requirements, the considered 360° video flows require between 16.5 Mb/s (everest, VMAF=70) and 78 Mb/s (buffaloes, VMAF=92). In what follows, we analyze and quantify how many of these flows can be supported in a 5G cell.

IV. Capacity study

A. 5G cell capacity model

In this section, we perform a numerical analysis to evaluate how 5G networks in different configurations can support VR streaming. To obtain the expected capacity of 5G networks we consider the User Equipment (UE) model specified in [13], which allows us to devise the maximum capacity depending on various parameters, such as the system bandwidth, the Modulation and Coding Scheme (MCS), the number of MIMO layers, etc. Besides, we cross-validated our results using the Simu5G [14] simulator.

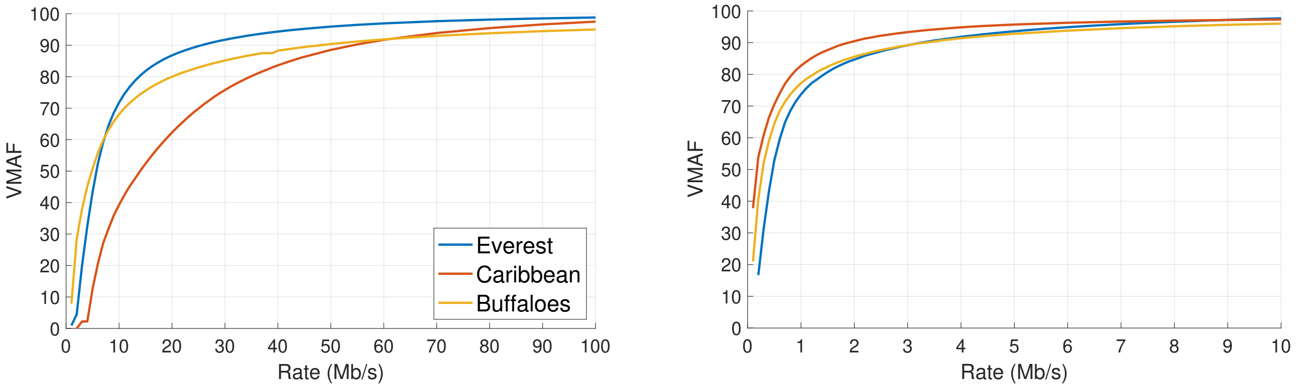
B. Impact of the MCS

Here we assume a bandwidth of 100 MHz, 4x4 MIMO, and numerology 1. As discussed above, we keep VMAF=95 for the front view and consider two profiles: $i)$ VMAF=92, where the full video is transmitted at the highest considered quality for this flow, and $ii)$ VMAF=70, where the full video is transmitted at the lowest considered quality. We consider all the available MCS and compute for each MCS, video, and profile the maximum number of streams that could fit in a 5G cell. We depict the results in Fig. 3.

The results show that the cell capacity increases with the MCS, with a steep increase for MCS ≥ 15, in particular for the VMAF=70 flows (indicated by dashed lines). Focusing on the VMAF=92 flows (solid lines), the capacity for transmitting the Caribbean and the Buffaloes videos is nearly identical, with both achieving a maximum of 22 simultaneous flows for the highest MCS. This number increases to 46 flows for the Everest video, which is caused by its lower demands in terms of throughput.

These findings underscore the substantial impact that the type of video has on the resulting capacity. This disparity is more pronounced with VMAF=70 videos. For the most demanding video (Caribbean) fewer than 60 flows can be

[2] Our methodology assumes that the downloaded AV1 videos are of very high quality. We have confirmed that the resulting throughput figures for the videos we analyzed match those reported in [12] for the transmission of AV1 video with VMAF=95, and therefore we have no reasons to believe otherwise.



(a) Equirectangular videos.

(b) Front view videos.

Fig. 2: Comparison of Equirectangular (left) and Front view (right) videos' VMAF vs encoding rate.

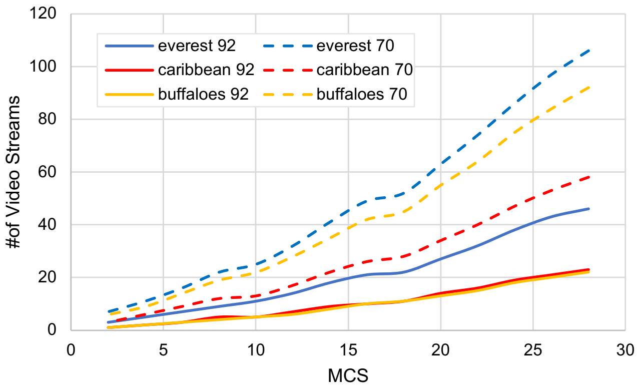


Fig. 3: Impact of the used MCS on the maximum number of video streams.

accommodated, while for the other two videos the capacity escalates to as many as 92 flows (Buffaloes) or 105 (Everest), highlighting the variability in demand across different video types.

C. Impact of the full video quality

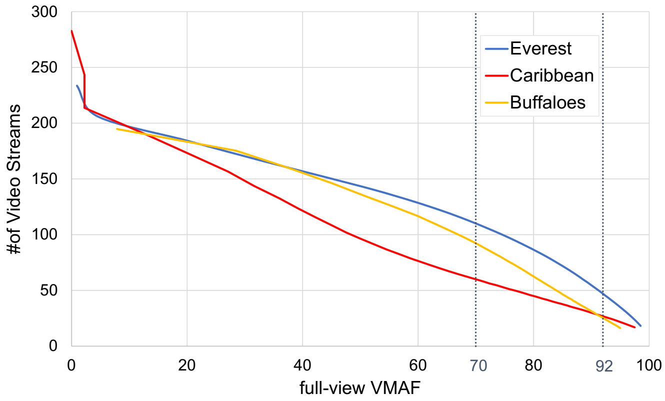


Fig. 4: Trade-off between cell capacity and QoE.

We next explore in more detail the impact of the quality of the full view on the capacity. To this aim, we assume the same cell configuration as before and assume the use of the highest MCS. As before, we keep the front view to VMAF=95 and perform a sweep on the VMAF of the full view, from 0 (i.e., no transmission) to the highest possible value. For each considered VMAF and video, we compute the maximum number of flows supported. We plot the resulting pairs of (VMAF, # streams) in Fig. 4, to illustrates the trade-off between the quality of the full view video (x-axis) and the maximum capacity (y-axis) (the highlighted values at VMAF=70 and VMAF=92 correspond to the same results at MCS=28 in the previous Fig. 3).

Firstly, it is worth highlighting the significant decline in the capacity as soon as the VMAF exceeds a marginal value, illustrated by the reduction of supported Caribbean video streams from approx. 275 to 225. This steep decrease in capacity can be seen as the *cost* of transmitting video outside the front view, which could be eliminated with precise and timely FoV prediction algorithms (so only the front view would be transmitted). Secondly, the figure illustrates an approx. linear inverse relation between the number of supported streams and the QoE, with a reduction of roughly 2 flows per each VMAF score increase. Third and finally, it is also worth noting again the impact of the type of video on performance, as e.g. at VMAF=60 the maximum capacity varies between 60 (Caribbean) and 130 (Everest), i.e., a span of 70 streams.

D. Non-ideal channel conditions

Finally, we explore the resulting capacity under more constrained transmission conditions with the use of `Simu5G`. We assume a scenario with four cells: one *tagged* gNB supporting the downlink transmission of VR flows which is surrounded by 3 interfering gNBs with an inter-site distance of 600 m. Each gNB has a bandwidth of 100 MHz and employs numerology 1, but in this case, no MIMO is used since it is not supported yet by the simulation tool –so the maximum achievable capacity is divided by 1/4 as compared to the previous case.

A variable number of User Equipments (UEs) is randomly deployed in the coverage area of the tagged cell, and moves within the coverage area at a constant speed of 50 km/h. For each of the videos defined in Section III and used in Section IV-B, we extract a trace file, wherein each video

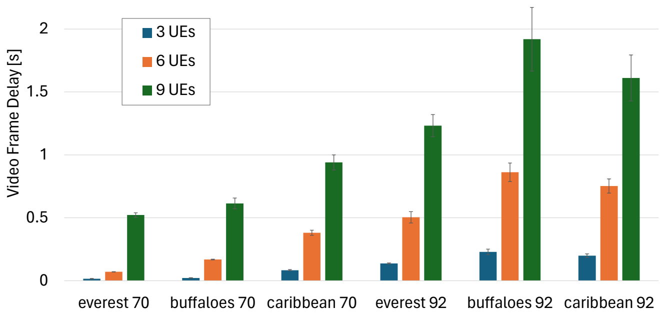


Fig. 5: Video frame delay for two video configurations of the full view (VMAF={70,92}) and an increasing number of UEs.

frame is associated with a timestamp and a size in bytes, and we let UEs in each configuration receive a UDP flow resembling the trace file of the corresponding video profile. For each considered number of UEs, video profile, and video, we compute the packet delay of each flow towards each of the UE, which is defined as the time between the transmission of a packet from the gNB and its reception at the UE. To analyze the system at maximum utilization, we configure the gNB to use a "max throughput" scheduler that gives priority to those UEs using the highest MCS. Although this could starve those UEs using lower MCS, the overall performance is better than with a "proportional fair" scheduler, as our results confirm (not presented due to space reasons). Each simulation is repeated 10 times to achieve statistical soundness, and confidence intervals at 95% level are reported.

We consider each of the three representative videos separately. For each video and quality, we compute the average delay from all UEs. We provide in Fig. 5 the results, where the bars represent the average delay, colors indicate the number of UEs considered, and results are grouped by video and quality profile. The results show that performance worsens with the number of flows, ranging between 0.5 s and 2 s for just 9 UEs (green bars). In fact, the highest quality profile (VMAF=92) already results very demanding with only 6 UEs, since the average delay exceeds 0.5 s for all videos considered. In contrast, with the VMAF=70 profile and 6 UEs the delay is kept below 0.4 s, which could be *acceptable* for some VR experiences. Finally, the results confirm that the "everest" video is the least demanding video in all considered cases, while the buffaloes video has smaller delays than the caribbean video at VMAF=70 (less traffic) but larger delays at VMAF=92 (more traffic), a result caused by its larger variability of the instantaneous transmission rate. Overall, these results confirm the need to use transmission profiles with moderate consumption of resources to maximize the capacity of the cell.

E. Main takeaways

In our study, we have illustrated that the maximum number of flows that can be supported in a 5G cell heavily depends on the variables of the scenario. We next summarize the main takeaways from our study: (1) under ideal conditions, there can be differences of 2× in the number of videos that can be admitted at the best quality in a 100 MHz, 4x4 MIMO 5G cell using numerology 1, i.e., from approx. 22 flows (caribbean, buffalos) to 46 flows (everest), as illustrated in Fig. 3; (2) if the quality of the non-FoV stream is reduced (Fig. 4), this capacity could be improved up to a factor of 5×, which illustrates the potential capacity gains of using accurate FoV-prediction algorithms that do not require a non FoV stream to deal with sudden head movements; (3) non-ideal channel conditions, instantaneous bitrate variations, and mobility can reduce the number of flows by a factor of 0.5×, according to the results from the `Simu5G` simulator (Fig. 5).

V. Summary and Future Work

In this paper, we have presented a methodology to analyze the capacity of 5G networks to support 360° video transmissions, which is pivotal for immersive Virtual Reality (VR) experiences. Our study quantifies the trade-off between video quality, characterized with VMAF, and the number of concurrent VR streams that a 5G cell can support. Our methodology paves the way for several lines of future work, such as: the use of alternate and/or more sophisticated video transmission models, including other video codecs; the development of call admission control schemes, that limit the number of VR flows to ensure the QoE of existing flows (including non VR flows); the impact of other aspects on performance (FoV prediction, beamforming, etc.); the relation between KQI and the resulting QoE; or the validation of the methodology based on VMAF using Mean Opinion Scores tests with real users.

Acknowledgements

We thank M. A. Martínez López from YBVR for the initial discussions that partly motivated this work. This work has been partly funded by the European Union-NextGenerationEU through the UNICO 5G I+D SORUS project, by European Union's Horizon-JU-SNS-2022 Research and Innovation Programme Project TrialsNet (Grant Agreement No. 101095871), and by the Italian MUR in the framework of the CrossLab and the FoReLab projects (Departments of Excellence) and PRIN 2022 project TWINKLE. We have used ChatGPT for editing and grammar enhancements.

References

[1] Netflix, "VMAF: Video Multi-Method Assessment Fusion," https://github.com/Netflix/vmaf, 2024, gitHub repository.

[2] M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, and N. García, "Video Multimethod Assessment Fusion (VMAF) on 360VR Contents," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 1, pp. 22–31, 2020.

[3] O. S. Peñaherrera-Pulla, C. Baena, S. Fortes, E. Baena, and R. Barco, "KQI Assessment of VR Services: A Case Study on 360-Video Over 4G and 5G," *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 5366–5382, 2022.

[4] H. Strasburger and E. Pöppel, "Visual Field," *Clinical and Experimental Optometry*, vol. 57, no. 1, pp. 33–33, 1974. [Online]. Available: https://doi.org/10.1111/j.1444-0938.1974.tb02786.x

[5] 3GPP, "Virtual Reality (VR) media services over 3GPP," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 26.918, April 2022, version 17.0.0.

[6] X. Hou, J. Zhang, M. Budagavi, and S. Dey, "Head and Body Motion Prediction to Enable Mobile VR Experiences with Low Latency," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–7.

[7] F. Duanmu, E. Kurdoglu, S. A. Hosseini, Y. Liu, and Y. Wang, "Prioritized Buffer Control in Two-tier 360 Video Streaming," in *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, ser. VR/AR Network '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 13–18. [Online]. Available: https://doi.org/10.1145/3097895.3097898

[8] AirPano VR YouTube channel. [Online]. Available: https://www.youtube.com/channel/UCUSElbgKZpE4Xdh5aFWG-Ig

[9] F. Gringoli, P. Serrano, I. Ucar, N. Facchi, and A. Azcorra, "Experimental QoE Evaluation of Multicast Video Delivery over IEEE 802.11aa WLANs," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2549–2561, 2019.

[10] J. Ozer. (2018, 11) Best Practices for Netflix's VMAF Metric. [Online]. Available: https://streaminglearningcenter.com/encoding/best-practices-for-netflixs-vmaf-metric.html

[11] ——. (2022, 5) Identifying the Top Rung of a Bitrate Ladder. [Online]. Available: https://ottverse.com/top-rung-of-encoding-bitrate-ladder-abr-video-streaming/

[12] ——. Estimating the Bitrate for 8K Videos When Encoding with HEVC and AV1 . [Online]. Available: https://streaminglearningcenter.com/ffmpeg/estimating-the-bitrate-for-8k-videos.html/

[13] 3GPP, "User Equipment (UE) radio access capabilities," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.306, September 2023, version 17.6.0.

[14] G. Nardini, D. Sabella, G. Stea, P. Thakkar, and A. Virdis, "Simu5G–An OMNeT++ Library for End-to-End Performance Evaluation of 5G Networks," *IEEE Access*, vol. 8, pp. 181 176–181 191, 2020.

Pablo Serrano (M'09, SM'16) is an Associate Professor at the University Carlos III de Madrid. He has over 100 scientific papers in peer-reviewed international journals and conferences. He currently serves as Editor for IEEE Open Journal of the Communication Society

Antonio Virdis is Senior Assistant Professor at the University of Pisa. He has over 80 peer-reviewed papers on the topics of Quality of Service, Edge Computing, network simulation and performance evaluation.

Francesco Gringoli is Full Professor at the University of Brescia, Italy. He received the Ph.D. degree in Information Engineering from the University of Brescia, Italy, in 2002. His research interests include security assessment, performance evaluation and medium access control in Wireless LANs.

Marco Gramaglia is a Visiting Professor at the University Carlos III of Madrid, where he received M.Sc (2009) and Ph.D (2012) degrees in Telematics Engineering.